

Biomedical boom, forget the rest?

As the US budget looms, the repercussions that President Bush's campaign promises will have on research and on other government programmes are becoming all too clear. Researchers face a tough struggle in lobbying Congress.

George W. Bush, the new president of the United States, is clearly determined to present himself as a man who means what he says. It should come as no surprise, then, that the hurriedly prepared 2002 budget that he is obliged by law to propose to Congress before the end of February will be shaped almost entirely by the direct pledges made during his election campaign. These included a promise to continue with plans to double the budget of the National Institutes of Health (NIH) over a five-year period, to boost investment in education, to balance the budget and to have enough left over to give Americans a significant cut in income tax.

Bush is likely to announce a \$3 billion increase in funding for the NIH up front, instead of leaving it to Congress, as his predecessors have done, to come up with extra money for its favourite research agency. This honest and straightforward move might be expected to thrill the scientific community, but it raises problems. The White House, unlike Congress, tends to consider the research budget as a whole in drawing up spending plans: NIH's hefty increase will leave almost nothing for other research agencies. The result will be a grotesque imbalance between spending in biomedical research and that in other disciplines of science. Rectifying this will now fall to Congress, where support for non-health research has traditionally been weakest.

Meanwhile, Bush, in order to assure the credibility of his tax-cut plan, will have to propose budget cuts in other government programmes — including social welfare programmes — to permit NIH's expansion. The danger is that the closure of nursery schools and homeless shelters to pay for more grandiose academic medical centres might fracture the bipartisan consensus that has driven the NIH expansion in Congress.

Even so, the \$3 billion increase at NIH is expected to stick, at least this year, leaving the rest of American science behind in the dust. The National Science Foundation, for example, may find its budget frozen, despite all the fine words about its role in underpinning scientific progress for every application, including health and defence.

The situation at the Department of Energy, which supports most US physics and runs many of the nation's large scientific facilities, is even bleaker. In the land of opportunity for all, physics budgets that have been flat for several years spell a dangerous stagnation for US physics. Since the last Bush left office, practically no major new facilities have been planned, built or even seriously contemplated.

Physical scientists' response to their predicament thus far has been to consider a new lobby shop in Washington that might emulate the recent successes of their biomedical brethren through groups such as Research!America. If the precedent of the Reagan administration is any guide, Washington will soon enjoy an influx of such lobbyists, representing every interest in the land that feels threatened by Bush's effort to control public spending.

But science needs to send a more sophisticated message than the railroad operator, the sugar-cane grower, the banana importer or the small restaurateur. A more intelligent response by the community would be to start facing facts, and to offer to help the administration set scientific priorities. That will require strengthening the science office in the White House to coordinate research policy and, in effect, prevent the sort of imbalance that will crop up in next week's budget proposals.

Bush will appoint his science team later this spring. They will soon learn whether the community is united behind such an approach — or bitterly divided between the haves and the have-nots. ■

Europe's infrastructure failure

Despite expectations to the contrary, the European Commission has failed to back research facilities for the long term.

European proponents of the much-heralded post-genomics revolution are suffering the curse of Sisyphus. They seem condemned for ever to gain support for European-level research infrastructures — and then to watch as politicians tip the plans, like Sisyphus's boulder, back to the bottom of the hill.

Two facilities, the European Mouse Mutant Archive and the European Bioinformatics Institute, are about to witness this perverse manoeuvre yet again. Everyone — politicians, bureaucrats and scientists — acknowledges the fundamental importance of these facilities, created with the help of the commission's fourth Framework programme of research (FP4). Yet a sustained commitment to keep them going — let alone to help them keep pace with genomics — has failed to materialize. After appearing to be convinced, the commission has backed down at crucial moments.

Consider the 1999 launch of FP5, which unexpectedly declared that "routine" services could not be funded (see *Nature* 402, 4; 1999). This was a response to the political aversion of European Union member states to giving long-term support for service facilities. No other

organization came to the rescue. Widespread outrage forced the commission last year to reallocate a small part of Framework money for the infrastructures using a new — and temporary — definition of "routine". This money has just started to be allocated (see page 967).

Since his appointment in 1999, research commissioner Philippe Busquin has been sympathetic to the problem of infrastructures. And he has shown true scientific vision, appearing to embrace a desire for a rational exploitation of the human genome sequence in Europe. The implications were that the research infrastructures needed for this would be resolved in FP6, due to be launched next year.

It now seems that FP6 will not deliver the goods after all. A draft of the commission's first proposal for the programme indicates that the basis for commission support will remain essentially unchanged: no funds for routine upkeep of established facilities, even though total money for infrastructures will increase.

Member states can change this if they want, but national interests tend to dominate their thinking, and a unanimous voice is unlikely to be heard. And researchers? Like Sisyphus, yet again. ■





Aftershocks

India opens its doors to international earthquake research
p968



Divided fortunes

Bush budget to boost bioscience but freeze physics
p969



Hot stuff

No let-up in Mount Oyama's outpouring of toxic gases
p970



Italian protest

Grandees of science call for more money, fewer restrictions
p972

European Union moves to curb moratorium on transgenic plants

Quirin Schiermeier, Munich

In a bid to end the impasse that has effectively prevented transgenic crops from being grown on European farms, the European Union (EU) has adopted tighter rules for their use.

The biotechnology industry welcomed the new rules, which it hopes will eventually lead to the acceptance of genetically modified (GM) crops in Europe. Environmental groups were more circumspect, indicating that they would accept the rules as long as the ban on the technology was maintained.

But in a move that highlights the limits of the EU's jurisdiction, six states issued a statement saying they would keep the ban in place.

Following last week's ruling by the governing council, the 15 EU member states have 18 months to implement the rules. In the meantime, the present *de facto* moratorium on the commercial planting of new transgenic crop varieties is likely to continue.

Under the new rules, suppliers of transgenic crop seeds must carry out strict assessments of environmental risk before the seeds can be licensed for use. If approved, the GM plants must be continually monitored for their environmental and health effects.

Local authorities and the public will be



No hiding place: new rules could open the way for transgenic tests of crops such as oilseed rape (above).

informed about which farms are growing the plants. Details of the experimental use of GM plants will also be made public. The plants will initially be approved for use for a maximum period of ten years, after which a renewed application must be made.

The directive also addresses concerns that GM organisms may spread resistance to antibiotics. Organisms containing antibiotic-resistance genes, widely used as genetic markers, must be phased out by 2004 for GM

plants put on the market and by 2008 for GM organisms used experimentally.

The directive was adopted after being approved last week by the European Parliament. But the nations behind the *de facto* moratorium — France, Italy, Greece, Denmark, Austria and Luxembourg — reaffirmed their intention to block GM crops altogether.

The six states called for extra rules on the traceability and labelling of GM products, ▶

Wheels start to turn for mutant mouse archive

Alison Abbott, Munich

After a somewhat shaky start, the European Mouse Mutant Archive (EMMA) is at last preparing to expand. With a grant from the European Commission, the archive is set to become an important source of mouse models for biologists.

The 500,000 euros (US\$460,000) awarded to the archive by the commission will allow EMMA to improve its organization and start a new database of mouse strains.

EMMA is a collaboration between its main site at Monterotondo near Rome and branches in Germany, France, Sweden, Britain and Portugal. The new database will be housed at the European Bioinformatics



Broker: Hrabé de Angelis.

Institute near Cambridge.

It will interface with a database at the Jackson Laboratory in Maine, currently the main source of mutant mouse strains worldwide. EMMA's archive is available free to all academics, and the mice cost 200 euros per strain.

EMMA has had a difficult birth, typifying the problems that can beset cross-European ventures. It struggled to open at all in 1999, after failing to win sufficient support from either national governments or the European

Commission (see *Nature* 402, 4; 1999). So far it has archived only about 100 mouse strains.

But its star appears to be rising following the appointment last autumn of its first director, Martin Hrabé de Angelis, of the Institute for Experimental Genetics at the GSF national research centre in Munich. Hrabé de Angelis has brokered an agreement between EMMA's sites to archive another 650 mouse strains over the next three years.

But EMMA's long-term funding is uncertain. The European Commission's sixth Framework programme for research has so far made no commitment to funding for existing research infrastructures. ■

▶ www.emma.rm.cnr.it

and they also want the commission to set up a system concerning environmental liability to supplement the existing regulatory framework for plant biotechnology.

The industry gave the directive a cautious welcome, despite its rigour. "It will provide the necessary transparency to restore public confidence in the safety of GM products," says Simon Barber, director of the plant biotechnology unit at EuropaBio, the European Association for Bioindustries. "This is an important step towards the end of the moratorium and it will hopefully enable us eventually to sell our products in Europe."

"We understand the demand from society for greater rigour," says Barber. But he warns that over-regulation of plant biotechnology could hinder innovation and put plant geneticists in Europe at a competitive disadvantage.

The moratorium has held up 14 applications for the commercial cultivation of GM plants, including maize and oilseed rape varieties developed by Aventis Crop Science, Syngenta and Plant Genetic Systems.

The environmental group Greenpeace called the directive "a massive improvement" on existing regulations. But a spokesman for the group said that it would not justify lifting the moratorium. ■

Canada's BSE fears 'groundless'

David Spurgeon, Montreal

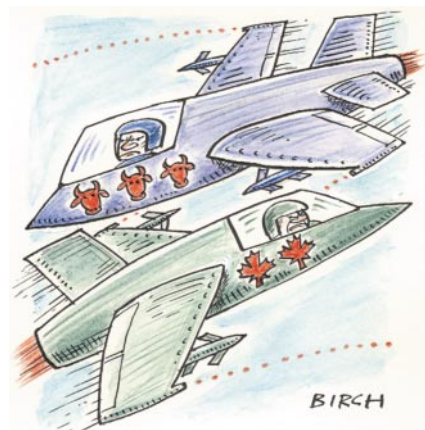
Brazil and Canada moved quickly this week to patch up a row over a Canadian threat to ban Brazilian beef because of fears of bovine spongiform encephalopathy (BSE).

Canada had announced that it would ban imports of Brazilian beef on scientific grounds. But after a barrage of criticism from both countries, the government indicated that the ban was likely to be lifted as suddenly as it had been imposed.

The threat to ban the beef caused uproar in Brazil, and came under fire in Canada after two scientists in the federal health department said publicly that it had no scientific basis and was motivated by politics rather than health concerns.

The Canadian Food Inspection Agency says that the ban was imposed because Brazil failed to supply adequate documentation on cattle it had recently imported from Europe. But critics say the real reason was a long-running trade dispute between the two countries over aircraft manufacturing subsidies.

The World Trade Organization has ruled that Brazilian subsidies to Embraer, a direct competitor of Canada's Bombardier in the lucrative market for small passenger-jets, are



illegal, and that Canada can impose retaliatory trade sanctions. But it has been reluctant to do this, as it regards Brazil as an important regional ally.

In Brazil, where no cases of BSE have been reported, the beef ban was branded a dirty trick in the press.

The mild-mannered Brazilian president, Fernando Henrique Cardoso, warned of a possible trade war with Canada, and the ban sparked public protests and led to Canadian whisky being dumped in the streets of São Paulo. ■

India promises more earthquake research

K. S. Jayaraman, New Delhi

The earthquake that ravaged Gujarat last month has drawn a pledge from the Indian government to increase seismological research and allow participation by foreign scientists.

India has been deeply suspicious of foreign involvement in its earthquake research. Apart from concerns that foreign researchers would use their data to detect Indian nuclear-weapons tests, India had alleged that US researchers abused their access to the Himalayas by attempting to spy on Chinese missile programmes.

And India has previously allowed foreign seismologists to carry out field studies only in collaboration with local scientists. "We don't want to allow outsiders to walk away with data without our people knowing what was collected, and for what purpose," explains Valangiman Ramamurthi, the science secretary.

But after a meeting of India's top scientific administrators, a change of approach was announced last week.

The Himalayas are of great interest to seismologists because they were created by the collision of the Indian and Eurasian tectonic plates, which resulted in the formation



Seismic shock: the Indian government plans to build 40 seismic monitoring stations.

of some two dozen active faults. The Rann of Kutchh, which on 26 January produced the country's worst earthquake in decades, is also of geological interest. But both regions are militarily sensitive.

Access by Western scientists was restricted after an incident in 1974 when it was discovered that a US expedition had left a nuclear-powered sensor on a Himalayan peak so as to spy on Chinese missile launches.

And in 1988, a US gravity survey of the Himalayas was halted after the Indian defence ministry said the data could be used to programme the path of US ballistic missiles flying over the mountain range.

Indian scientists have also been refused access to parts of the Rann of Kutchh — including the vast Allah Bund (wall of God) scarp created by an earthquake in 1819 — because they lie on the heavily militarized border with Pakistan.

Vinod Gaur, a seismologist with the Centre for Mathematical Modelling and Computer Simulation in Bangalore, welcomed the announcement. "I don't underestimate the security risks near the border, but our attitude should be to encourage important scientific investigations," he says.

Even before the announcement, the Gujarat earthquake seemed to have eased the clampdown on foreign seismologists. Ramamurthi says that more than 30 foreign groups are already studying the aftershocks and other aspects of the earthquake.

The government has also announced a US\$12 million plan to build 40 seismic monitoring stations in the Himalayas, where it fears the next major earthquake will occur. ■

First Bush budget set to favour life sciences

Matthew Davis, Washington

The first budget proposal of US president George W. Bush is expected to seek billions of additional dollars for biomedical research — but precious little for any other scientific programmes.

Sources familiar with the plan, which will be released early next week, say Bush will ask for a \$3 billion (15%) boost for the National Institutes of Health (NIH), making good on his campaign promise to carry on the effort begun three years ago to double the agency's budget by 2003.

But Bush is likely to seek an increase of just 1–2% for the National Science Foundation (NSF) — the main backer of non-biomedical university research in the United States — which last year got a 14% increase. He is expected to maintain funding for science programmes at the Department of Energy, which supports most US physics, at current levels and may slash the US Geological Survey's \$885 million budget by 20% or more.

A Republican congressional official familiar with the Bush administration strategy says that the numbers do not reflect any hostility towards the physical sciences. Rather, the embryonic administration, which has had to frame its budget before many of its senior positions have been filled, is trying to fulfil specific campaign promises. That means increasing spending for education, the NIH and perhaps defence, while controlling the overall budget sufficiently to allow for a promised cut in income tax.

Research administrators in the physical sciences are alarmed at the possible impact of this strategy on their budgets. They fear that, despite their recent efforts to make a case for a fairer balance between disciplines, the outcome will deepen what they see as an imbalance in favour of biomedical research.

At a recent meeting with senior White House budget official Mitch Daniels, Sherwood Boehlert (Republican, New York), who chairs the House Science Committee, complained about Bush's meagre allocations for research at the NSF, NASA and other agencies. According to one source at the meeting, Daniels displayed a detailed knowledge of the various science programmes. But he left a clear impression that Bush's proposed spending increases will remain tightly focused on initiatives, such as doubling the NIH budget, that were the subject of campaign promises.

Some NIH advocates are nervous that the agency's unprecedented expansion will come into direct conflict with other programmes for the first time. They fear that if other health and social-assistance programmes are cut to allow the NIH to grow, the broad coalition that has previously supported the agency in Congress may fracture.



Senate budget chair Pete Domenici (left) and others in Congress may augment Bush's spending plans.

But last week NIH supporters drummed up support for a non-binding resolution introduced by Senator Arlen Specter (Republican, Pennsylvania) and Senator Tom Harkin (Democrat, Iowa) that would put the Senate on record as supporting an extra \$3.4 billion for the NIH in 2002.

With Specter and Harkin leading the charge, Congress has handed the NIH an additional \$6.7 billion over the past three years, with no other programmes appearing to have suffered as a result. That was because a mounting budget allowed President Bill Clinton and the Congress to engage in a succession of spending sprees that satisfied almost every constituency.

Things could be different this year. Budget

experts on Capitol Hill say that, with Republicans controlling both Congress and the White House, and tax cuts and a slowing economy reducing the surplus, the days of unbridled appropriations — even for a Republican favourite like the NIH — may be over.

And at less fashionable agencies such as the Department of Energy, which was initially told by the new administration to prepare for a \$1 billion reduction in its \$18 billion budget, administrators fear that merely maintaining current funding could lead to stagnation for the physical sciences. Their best hope is that Congress, which will consider the Bush request this summer before passing a final budget by October, will be persuaded to augment the Bush proposal. ■

Farming accused of eco-damage

Corie Lok, Washington

The growing intensity of agriculture around the world is destroying soil quality, wasting water, demolishing forests and adding to greenhouse-gas emissions, according to a report by the International Food Policy Research Institute (IFPRI) and the World Resources Institute (WRI).

The *Pilot Analysis of Global Ecosystems: Agroecosystems* is billed by its authors as the first comprehensive audit of agriculture's global ecological impact. It argues that as much as four-fifths of the world's farm land is being degraded by agricultural effects such as erosion, nutrient depletion, acidification and loss of organic matter.

The audit makes a good case for more investment in agricultural research, especially in the developing world, says Ian Johnson, chair of the Consultative Group on International Agricultural Research and a vice-president of the World Bank. The bank is expected to use the report in its own

forthcoming assessment of global ecosystems.

Agricultural science and technology, including genetically modified crops and plant breeding, are critical to raising food production while protecting natural resources, adds Robert Thompson, director of rural development at the bank.

The report says that farmers withdraw 70% of the world's fresh water supply for irrigation, but more than half of this is lost through evaporation. Farm land has grown by around 130,000 square kilometres a year over the past 20 years, mostly at the expense of forests and grasslands, it says, and farming is also encroaching on many national parks and other protected areas.

The audit also finds that agriculture is the largest source of methane — a potent greenhouse gas — from human activities, responsible for 44% of all emissions. And it calls for more investment in the science of ecosystem monitoring. ■

www.ifpri.org

Residents wait for return to Japanese volcano island

David Cyranoski, Tokyo

Six months after its main eruption, Mount Oyama on the island of Miyake, 200 kilometres south of Tokyo, is still belching thousands of tonnes of toxic gas into the atmosphere every day. And Japanese volcanologists say they do not know when it is likely to stop.

This month, a meeting of the Coordinating Committee for the Prediction of Volcanic Eruptions admitted that there is no end in sight for the sulphur dioxide (SO₂) emissions.

The extraordinary volume of SO₂ being given off makes it advisable to keep the residents off the island, the scientists say. Miyake's 3,900 residents have been evacuated to the mainland where health risks from the gas emissions are said to be slight.

Oyama drew attention late last June when researchers detected a movement of rising magma, which resulted in a violent underwater eruption to the west of the island. In August, a series of small earthquakes and eruptions led to the main blast on 18 August, which shot volcanic smoke several kilometres into the air and created a mammoth 1.5-kilometre crater at the volcano's mouth.

There have been no strong tremors since August, leading scientists to conclude that there is little reason to expect further explosive eruptions. But they point out that the temperature at the volcano's mouth peaked in December and remains high, indicating that there is still a risk of small eruptions.

But SO₂ continues to pour into the sky at a rate of tens of thousands of tonnes per day.

"This long-term expulsion of gas has never really been observed before," says Yoshiaki Ida, professor of geophysics at the University of Tokyo and chair of the coordinating committee.

The emissions are not the only surprise at Miyake. Because Oyama is a 'basaltic' volcano, it should generate a steady emission of fluid magma rock, rather than producing the violent eruptions that have characterized this incident. The fact that an eruption on the edge of the mountain, away from the volcanic peak, preceded the main eruption is also abnormal. "Unusual explosions must have been triggered by the preceding flank eruption and the collapse of the summit area, but the precise mechanism is unknown," Ida says.

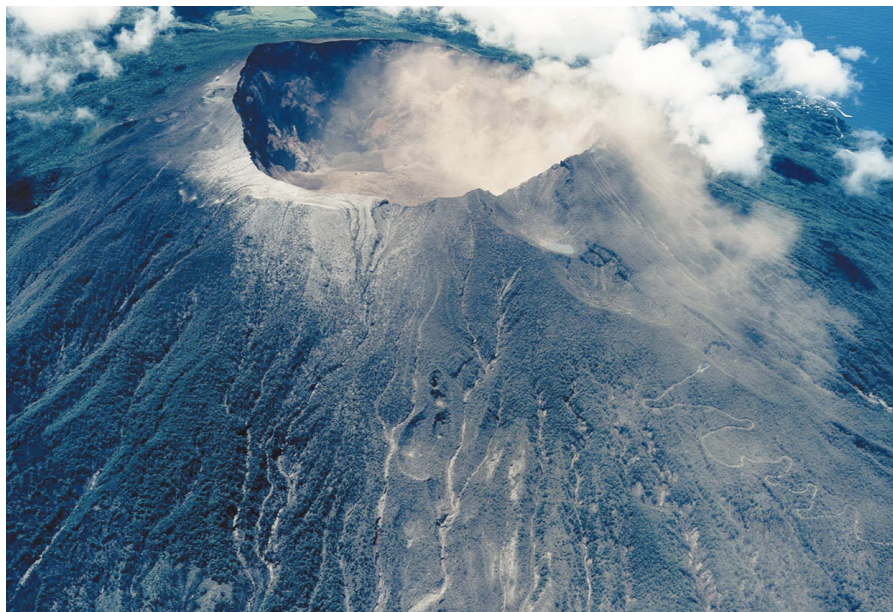
Japanese volcanologists will continue to study these anomalies, and try to figure out the reasons for them. "It could be that this long-term expulsion of gas is related to the large mouth size of the crater," says Ida.

In the meantime, the island's displaced inhabitants await permission from the Tokyo metropolitan Disaster Prevention Planning Section — which takes advice from Ida's committee — to return home.

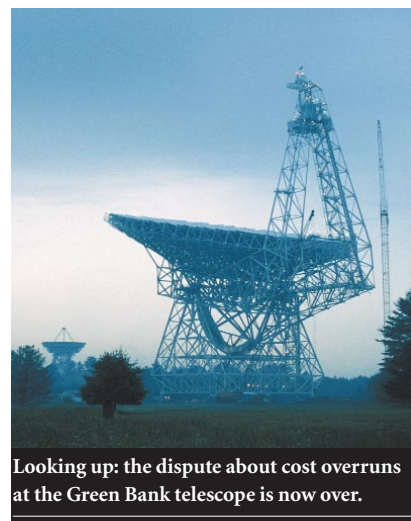
The committee will next meet in May — but residents could be waiting much longer than that. "A preliminary estimate of [volume of magma in the volcano] shows that the volcano has capacity to supply sulphur dioxide for several years, but the emission may end earlier than that," says Ida.

♦ <http://hakone.eri.u-tokyo.ac.jp/vrc/erup/miyake.html>

♦ <http://www.aist.go.jp/GSJJ/~kaz/miyakegas/COSPEC.html>



Blowing its top: Mount Oyama spews out thousands of tonnes of toxic gas every day.



Looking up: the dispute about cost overruns at the Green Bank telescope is now over.

Small payout over Green Bank telescope cheers astronomers

Irwin Goodwin, Washington

An arbitrator has ruled that the government should pay only \$4 million towards cost overruns encountered by the contractor that built a huge radiotelescope at Green Bank, West Virginia.

The decision — in the face of claims from the contractor for \$29 million on top of its \$55 million contract to build the world's largest fully steerable, single-dish telescope — was good news for US astronomers, who feared the settlement would eat into the astronomy budget at the National Science Foundation (NSF).

The contractor, Radiation Systems Incorporated — later part of the Comsat Corporation — began work in 1990, with completion planned for 1994. But the telescope achieved first light on 22 August 2000 and was accepted by the National Radio Astronomy Observatory (NRAO), on behalf of the NSF, on 13 October.

In 1998, Comsat sought a further \$29 million, arguing that NRAO and its managing contractor, Associated Universities Incorporated (AUI), had forced it to perform extra work.

"This matter is finally resolved, and we can now focus our efforts on making Green Bank a world-class telescope," says Paul Vanden Bout, NRAO's director.

Three days after it achieved first light, the GBT was opened by its original sponsor, senator Robert Byrd (Democrat, West Virginia), whose name it now carries.

But even as AUI and NSF officials worried about the possibility of an expensive settlement, Byrd discreetly slipped \$15 million into the NSF's budget bill for this year to cover astronomy facility costs, including \$8 million specifically for the GBT.

Biologists urge US to build marine reserves

Mark Schroppe, San Francisco

Protected 'marine reserves' off the coasts of the United States received unequivocal support this week in a statement issued by 160 prominent marine scientists. The group declared that there is now evidence that these reserves are effective as conservation tools.

The statement was crafted by the National Center for Ecological Analysis and Synthesis at the University of California at Santa Barbara, and was released at the annual meeting of the American Association for the Advancement of Science in San Francisco. It was issued in response to indications that the US government is preparing to stall plans to establish a network of the reserves.

"The message is clear," says Jane Lubchenco, a marine biologist at Oregon State University and one of the statement's signatories. "Marine reserves work, they work quickly, and they are a powerful and underutilized tool. Existing scientific information justifies immediate application of networks of marine reserves as a central management tool," she says.

In May of last year, former president Bill Clinton issued an executive order requiring a study of marine protected areas, where some activities are prohibited. The executive order also called for the planning of a network of such areas, probably including a small area of

'marine reserves', where no resource removal of any kind is permitted.

But in January, James Hansen (Republican, Utah), the new chairman of the powerful resources committee in the House of Representatives, sent a letter to the incoming administration of President George W. Bush. In it, he urged Bush to slow down any movement towards a network of marine protected areas, due to what he termed insufficient knowledge on their effectiveness.

Lubchenco says that, although a few years ago this was the case, the present state of knowledge makes it clear that the marine reserves would work.

The latest statement was released in conjunction with the results from several marine reserve studies to be published in a forthcoming issue of *Ecological Applications*. One, led by Robert Warner of the University of California at Santa Barbara, examined data on 81 marine reserves around the world and found almost universal and dramatic increases in population densities, organism size and species diversity within the reserves.

Roger Griffis, policy adviser for the National Oceanic and Atmospheric Administration, says that fishermen are still concerned that reserves will harm their business. But he notes that reserves may actually prevent further declines in catches.

Griffis adds that the decline of marine



Economies of scale: marine reserves could reverse declines in fish stocks, researchers say.

resources has reached a critical point and that, although study of reserves should continue, the recent studies amply show their effectiveness. "We do know enough to be acting," he says. ■

Scientists seek solidarity in oaths

Paul Smaglik and Colin Macilwain, San Francisco

Should scientists, like medical doctors, pledge to do no harm? The question, which has largely lain dormant for years, was revisited this week by US researchers.

A symposium held at the annual meeting of the American Association for the Advancement of Science (AAAS) in San Francisco considered the merits of adapting the Hippocratic oath to encompass all scientific disciplines.

Proponents say such an oath would help to regain public trust. But opponents say that it might cut off avenues of scientific inquiry that may eventually benefit society.

The movement to adopt oaths peaked in the United States during the nuclear arms race. It has now resurfaced in Europe, largely as a result of public concern about the role of science in the bovine spongiform encephalopathy debacle and in genetically modified food.

French scientists are seriously considering the merits of scientific oaths, says Gerard Toulouse, a scientific director at the Laboratoire de Physique Théorique de l'École

Normale Supérieure in Paris. He says they might provide an impetus to make scientists more conscious of their public duty.

But Irving Lerch, head of international affairs for the American Physical Society, is not sure that oaths will win public trust. He says it is hard for scientists to pledge to work only for the public good when they can predict neither the outcome of their research nor how someone else might apply it. And to be of any use, he notes, an oath would have to be enforced by someone. Scientific societies seem reluctant to take on such a role.

But a few blocks away from the crowded AAAS meeting, a clutch of pressure groups and scientists called on researchers to sign a pledge right now, disavowing work on all weapons of mass destruction.

One pledge supporter, Greg Mello of the Los Alamos Study Group, says that societies such as the AAAS — which declined to accommodate the pledge press conference — have vague discussions about oaths, but "shy away from anything that might affect peoples' careers". ■

► <http://www.iasg.org/pledge>

Climate change offers bleak future

Quirin Schiermeier, Munich

Global warming is damaging natural systems across the whole planet, according to a report from the international group of scientists charged with studying climate change.

All continents will suffer economically, the report says, but Africa, Asia, South America and the small island states will be most affected.

The report is the second in a series of three from the Intergovernmental Panel on Climate Change (IPCC). Some natural systems, including glaciers and coral reefs, "may undergo significant and irreversible damage", the panel says.

In particular, the scientists predict a reduction in crop yields in warm countries, decreased water availability in regions where water is already scarce, an increase in flooding risks, and the spread of diseases such as malaria and cholera.

The "impacts of future changes in climate extremes are expected to fall disproportionately on the poor", they add. ■

AP

Italian scientists take to the streets against political interference

Rome More than 1,000 scientists, led by 91-year-old Nobel prize-winning neurologist Rita Levi-Montalcini and pharmacologist Silvio Garattini, marched in Rome last week to protest against low science budgets and restrictions on biology research in Italy.

The demonstrators expressed anger at what they see as ideological and religious factors dictating government decisions on contentious issues such as genetically modified food and stem-cell research, and about low Italian investment in research.

Much of the protesters' ire was targeted at Alfonso Pecoraro Scanio, the agriculture secretary and a member of the Green Party, for his opposition to experiments with genetically modified crops. But after a meeting with Pecoraro Scanio and Prime Minister Giuliano Amato, the scientists obtained a promise that a panel would look into such experiments.

In advance of this spring's elections, scientists have also submitted a questionnaire to Francesco Rutelli and Silvio Berlusconi, the respective leaders of the government and opposition coalitions,



Age of dissent: Rita Levi-Montalcini with a poster protesting at restrictions on science.

asking about their attitude to scientific issues, including research on transgenic plants and stem cells.

Radioactive mineral exhibits put museum in the dock

London The Natural History Museum was in legal trouble last week, when a court found that it had exposed visitors and staff to radioactive minerals.

Museum trustees admitted that hundreds of rocks on display contained higher levels of uranium and thorium, and released as much as 50 times more ionizing radiation, than the curators believed.

Relatively few of the museum's visitors went to look at the rocks, and even

enthusiasts viewing them every six months would receive doses of only 45 microsieverts each year — less than one-hundredth of the recommended maximum exposure.

The museum was ordered to pay £6,300 (US\$9,100) costs for the case, which was brought by the Health and Safety Executive.

Intensive fish farming 'destroys ecosystems'

Washington Intensive fish farming sometimes consumes more fish as fish meal than it produces, degrades coastal habitats, and spreads disease to wild fish stocks, says an Ecological Society of America report.

Aquaculture production has more than doubled in the past 15 years but harvesting of wild fish and shellfish has not subsided, increasing the pressure on fish and other aquatic populations, the report says.

The authors of the report, *Effects of Aquaculture on World Fish Supplies*, say that production of the ten most commonly farmed fish, including salmon and shrimp, processes an average of 1.9 kilograms of wild fish into fish meal for every kilogram of farmed fish produced. Aquaculture also disrupts the oceanic food chain, and results in the loss of coastal wetlands and the introduction of disease and excess nutrients from wasted fish meal, the authors say.

Indian genomics gets a billion-rupee boost

New Delhi The Indian Council of Medical Research has pledged a billion rupees (US\$21.5 million) over five years to support 70 research projects in genomics, building on the newly published human genome sequence (see *Nature* 409, 814–958; 2001).

The projects are aimed at developing drugs and vaccines for diseases including malaria, tuberculosis, hepatitis and AIDS, says council chief Nirmal Kumar Ganguly. C. P. Thakur, the health minister, says the government will also support new centres for bioinformatics, molecular medicine, structural biology and stem-cell research.

Quebec offers tax holiday to academics

Montreal The state government of Quebec is offering a five-year tax holiday to attract foreign science and engineering professors.

Camille Limoges, deputy minister for science and technology in the province, says that Quebec's universities need the boost because they were built in the 1960s and 1970s, and a whole generation of staff is now retiring. The tax exemption will apply to those hired under the Canada Research Chairs programme (see *Nature* 401,

731–732; 1999): 570 of the 2,000 chairs to be funded will be located in Quebec.

Candidates for the scheme, which is expected to cost less than Can\$3 million (US\$2 million) annually, must have a PhD and agree to stay in the province for at least five years. Quebec already offers comparable benefits for highly skilled staff in industry.

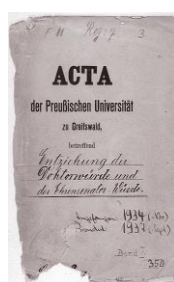
German university rehabilitates Nazi victims

Munich The University of Greifswald has formally rehabilitated 76 academics whose qualifications were removed by the Nazis.

Under a 1933 law, at least 1,600 academics at German universities — mainly Jews, communists, immigrants and academics

showing so-called 'antisocial behaviour' — lost their jobs. After the war, the universities were slow to rehabilitate the victims: Greifswald, for example, had rehabilitated only 12 of them by 1998.

But last year, after a recommendation by the Conference of German Educational Ministers, German universities



Act of malice: the official record of disqualifications.

agreed to search their archives for people whose academic titles were removed in the Nazi era.

Dirk Alvermann, head of the University of Greifswald's archives, says the university will continue its search. "We are now looking for university teachers who were hindered by the Nazis from their work, and whose careers were often stopped," says Alvermann.

Evolution allowed back into Kansas schools

San Diego Evolution and the geological formation of the Universe were restored to the curriculum in Kansas schools last week after a two-year battle with creationists, who briefly had the Bible's account of the beginning of the world taught instead.

The Kansas Board of Education approved new science standards that will direct schools from kindergarten to high school to teach evolution and the Big Bang theory of the development of the Universe. The change, approved by seven votes to three, was the result of last year's election of pro-evolution board members, backed by a state-wide network of scientists and their supporters.

Kansas is one of several states that have recently turned back attempts by religious fundamentalists to prevent the teaching of evolution and modern geology theory.

Let there be light

A silicon laser would revolutionize telecommunications, electronics and computing. Squeezing light out of silicon is no easy task, but Philip Ball discovers that researchers are becoming more optimistic about its light-emitting abilities.

The information age suffers from a split personality. Deep beneath the ocean's surface, photons of light stream through optical fibres, carrying voice and Internet traffic between continents. But before routing devices, computers and telephones can use the data, this light-borne information must be converted into electronic signals. With entirely optical computers unlikely to replace electronics in the near future, this uneasy marriage of electrons and photons is likely to persist for some time.

Improving the interface between silicon electronics and photonics is high on the agenda in the field of optoelectronics. Some researchers believe that the solution lies in reforming silicon's character. Given patience

and diligence, they say, the world's favourite semiconductor can be coaxed into emitting light. "If an all-silicon laser could be created, it would revolutionize the design of supercomputers and lead to new types of optoelectronic devices," says Leigh Canham of biomaterials company pSiMedica, a spin-off from the UK Defence Evaluation and Research Agency (DERA). And if recent progress continues, that revolution may not be far off.

Communication breakdown

At the moment, laser diodes are used to turn electronic signals into light pulses. These miniature lasers are built from layers of different semiconductors, named III-V alloys after the columns of the periodic table from which their constituents come. Positive and

negative charge carriers move into thin layers within the laser known as 'quantum wells'. Here the carriers recombine, releasing their energy as a photon of light.

But it is difficult to incorporate III-V alloys into silicon circuits. The two do not fit together because the spacing between the atoms in the two materials, known as the lattice constant, is different. The ideal laser diode for optical telecommunications — a blend of indium, gallium, arsenic and phosphorus, denoted $\text{In}_x\text{Ga}_{1-x}\text{As}_y\text{P}_{1-y}$ — illustrates the problem. This material emits light at a wavelength of about 1.5 micrometres, which is the optimum for transmission through glass optical fibres, but has a lattice constant that is 8% bigger than that of silicon.

This means that atoms at the interface

between the two materials do not match up, and line-like distortions form in the semiconductor. "Dislocations form near the interface and then thread through the III-V layer," explains Vincent Crespi, a physicist at Pennsylvania State University. Because this layer is thinner than the silicon substrate, the light-emitting III-V material incurs most of the deformation, and the resulting defects



Glowing future: Leigh Canham (left) and Vincent Crespi hope to make silicon optoelectronics a reality.

severely degrade the layer's electrical conductivity.

As a result, semiconductor laser diodes must be kept separate from silicon circuits, which hinders the constant drive within the electronics industry to reduce the size of its circuitry. Separate lasers introduce further problems when chips are connected together, because achieving and maintaining precise alignment between the chips and the light sources becomes more complicated. It would be far better if light-emitting devices could be integrated directly with silicon chips.

That would be simple if silicon itself emitted light. But normal 'bulk' silicon does not, for a subtle reason. Semiconductors emit light when their electrons jump between energy levels. The size of the energy gap between these states — known as the band gap — determines the wavelength of the photon emitted.

Silicon's band gap corresponds to light with a wavelength of 1 μm , which would be fine for fibre-optic transmission. But silicon has what solid-state physicists call an 'indirect' band gap, meaning that any electron moving between the two energy states must change its momentum. This makes the transition less likely to occur.

Porous potential

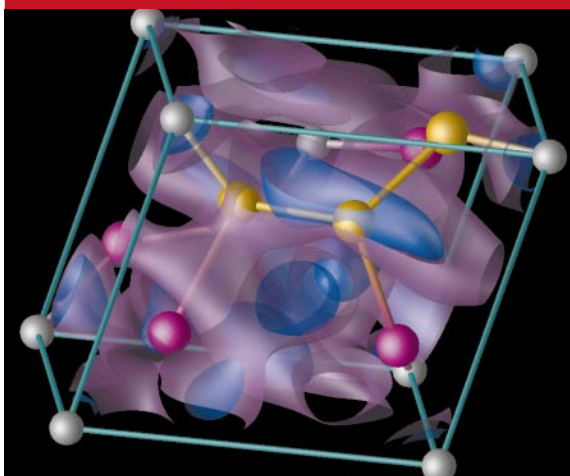
Despite these problems, researchers trying to persuade silicon to glow are currently in an upbeat mood. Silicon, they have found, does emit light if chopped down into tiny pieces just a few nanometres wide.

The first evidence for this remarkable change in silicon's behaviour emerged in 1990 from Canham's lab, then based at the Defence Research Agency (DERA's earlier incarnation) at Malvern in the English Midlands. Canham used hydrofluoric acid to dissolve silicon, etching away much of the material to leave a porous substance made up of a network of silicon 'nanowires'. These tiny threads confine mobile electrons to narrow channels. As the space available to the electrons shrinks, quantum 'confinement' effects come into play. These affect the band gap, making it much easier for the electrons to move between energy states.

The benefits are impressive. Illuminate silicon nanowires with laser light to create pairs of negative and positive charge carriers and suddenly porous silicon glows with visible light. Known as photoluminescence, this process is 10,000 times more efficient in the nanowires than in normal silicon.

Ordinarily, silicon would emit infrared light, but Canham reasoned that the band gap is increased by the quantum confinement effect. As a result, the wavelength of the emitted light gets shorter as the nanowires get thinner. Because increasing the etching time produces narrower wires, the system can be tuned to produce light of different colours. Silicon that emits red, orange or

Searching for silicon's perfect partner



This modelled blend of carbon, silicon and tin should emit light and match up with silicon's lattice spacing.

As squeezing light out of silicon has tended to bring the words 'blood' and 'stone' to mind, physicists are always on the lookout for new ways to integrate optical devices with silicon chips. This search has led William Gillin of Queen Mary College in London down an unusual path — he has gone organic.

Gillin's organic (carbon-based) materials sit happily on a silicon surface because they are amorphous, rather than crystalline — there is no lattice mismatch because there is no crystal lattice. But the trick is to find such light-emitting substances that are sufficiently conducting to function in electronic devices. Gillin and his colleagues have created a silicon-based organic light-emitting diode (OLED) using a layer of erbium tris(8-hydroxyquinoline) deposited on another organic compound⁸. Positive charge carriers from a

silicon substrate pass through the first organic layer to the erbium-containing layer, where they meet negative charge carriers provided by a top contact of aluminium. As the charge carriers recombine, they emit infrared radiation.

But these OLEDs have a long way to go before they are practical. The voltage needed to switch them on — 33 volts — is massive by industry standards, and their efficiency, just 0.01%, leaves a lot to be desired. Gillin is optimistic that it will be possible to get his devices to work at as little as 3 volts. They also have the advantage of being relatively easy to make and can emit light of different wavelengths if the erbium is replaced with different metals.

Other physicists are hoping to sidestep the problem of lattice mismatch by rational design. Rather than conducting trial-and-error searches for the right crystalline material to integrate

with silicon circuits, they are using computational methods to devise semiconductors with the desired properties — both in terms of lattice constant and the band gap (see main text) needed to produce light of useful wavelengths.

John Joannopoulos and colleagues at the Massachusetts Institute of Technology have calculated that a specific blend of zinc, silicon, phosphorus and arsenic should have a lattice constant only 0.08% smaller than that of silicon¹⁰, and should emit light at the ideal frequency for telecommunications. The only snag is that it is likely to be extremely hard to make.

Others have been searching for alloys that should be more amenable to synthesis. Vincent Crespi and his colleagues at Pennsylvania State University have teamed up with Arizona State University's John Kouvetakis, a materials scientist who specializes in synthesizing complex inorganic compounds. Crespi's team¹¹ has modelled blends of carbon, tin, germanium and silicon that have lattice constants within 1% of silicon's, and optimal wavelengths for fibre-optic transmission.

Kouvetakis is now trying to make these materials. So far, he has produced alloys of tin and germanium that are stable up to 400 °C and in which, crucially, the tin atoms seem to occupy the correct sites. Normally, tin separates out into clusters in these materials, destroying the crystal structure in the process. "The next challenge," says Crespi, "is to get the proper ratios of tin, silicon and germanium."

yellow light is reasonably straightforward to make. After this, things get tricky because the tiny wires become increasingly fragile. But porous silicon that emits green and blue light has been reported².

The prospects for silicon-based optoelectronic devices were boosted in 1996, when Philippe Fauchet of the University of Rochester in New York created a porous-silicon light-emitting diode (LED) integrated onto a chip³. LEDs are not bright enough for long-distance optical telecommunications, but they can be used for communicating over the short distances between and within

chips. Before porous silicon came along, researchers trying to integrate LEDs into chips had been hampered by the same problems affecting lasers: the incompatibility of materials.

Fauchet's silicon-based LED is a proof of concept, but it is not yet technologically viable. The energy efficiency and switching speeds — the rate at which the LED can be turned on and off — still need to be improved. But progress is being made.

Nobuyoshi Koshida and his colleagues at the Tokyo University of Agriculture and Technology have greatly increased the efficiency of

these LEDs. They used an oxidation reaction to eliminate much of the bulk-like 'non-confined' silicon that remains in the porous network — which otherwise transports most of the charge carriers⁴. Koshida's group has now raised the efficiency to about 1%, an improvement of five orders of magnitude over the earliest results. This is fine for display screens, says Canham. But to provide the photonic signals needed to transmit information between chips will require a further 10-fold increase. And switching speeds are still about two orders of magnitude too slow.

Dot comms

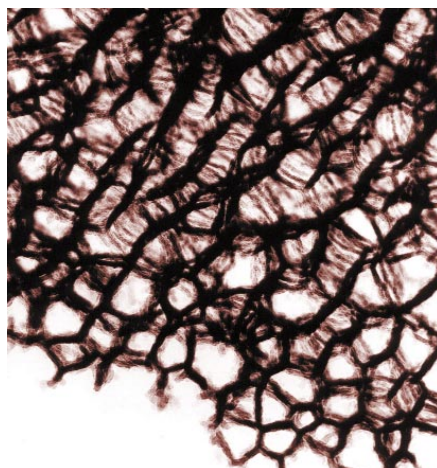
Using a network of silicon wires is not the only way to make the element glow. The same helpful quantum effects operate if the porous silicon is divided into nanometre-sized particles known as nanocrystals or 'quantum dots'. Last year, Munir Nayfeh of the University of Illinois at Urbana-Champaign and his colleagues used ultrasound to shatter porous silicon into nanocrystals. The smallest of these particles (about 1 nanometre across) emit blue light⁵.

Other teams have reported similar successes. In February last year, Brian Korgel of the University of Texas at Austin and his colleagues described how to grow single nanowires that emit blue light⁶. Unlike porous silicon, which consists of a mesh of nanowires, Korgel's nanowires are discrete threads of silicon just 4–5 nm wide.

Lorenzo Pavesi of the University of Trento in Italy has used an alternative approach to create silicon nanocrystals. By firing high-energy silicon ions into quartz (silicon dioxide), and then heating the material to 1,100 °C, Pavesi and his colleagues generated silicon particles about 3 nm across that were embedded in the quartz. Last November, these researchers showed that not only do the nanocrystals emit red light when energized with a laser beam, but they can also amplify a 'probe' beam of the same wavelength as the emission⁷. Known as optical gain, this phenomenon is one of the fundamental features of laser emission.

Although the Italian team has taken the first steps towards creating a silicon laser, light amplification is not the same as laser action. In a laser beam the light is coherent: all the photons are in phase. To achieve this, emitted photons must stimulate the emission of others — the stimulated photons emerge in step with those that induced them. This 'stimulated emission' is achieved by placing the emitting material in an optical cavity bounded by mirrors which let photons bounce back and forth.

Tantalizingly, at the Materials Research Society meeting in Boston last December, Nayfeh reported optical gain and stimulated emission from his blue-light-emitting nanocrystals. Others in the field are now waiting for Nayfeh to publish quantitative



Shining example: networks of porous silicon nanowires emit light in the visible spectrum.

data so that they can assess these claims.

But there might be another path to developing a silicon-based laser. When Ulf Gennser was working at the Paul Scherrer Institute in Villigen, Switzerland, he and his colleagues described a quantum-cascade laser (QCL) consisting of alternating layers of silicon and a germanium-silicon alloy⁸. These devices contain several five-layer blocks of light-emitting units, stacked one on top of the other. Electrons can 'tunnel' between individual layers, emitting a photon in the process.

Under the right conditions, this should lead to stimulated emission of coherent light, but the Swiss team has so far managed to produce only electroluminescence, not laser emission. "There remain many obstacles before we have a working laser," admits Gennser, who is now at the Laboratory of Microstructures and Microelectronics in



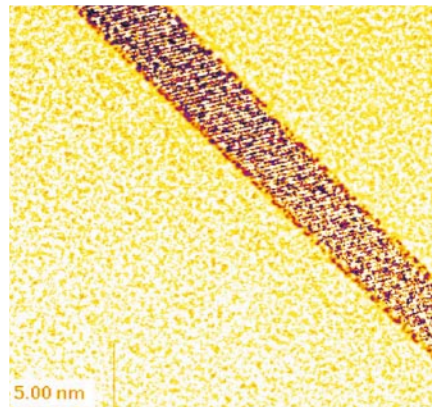
Light speed: single silicon nanowires (inset) might supply photonic signals to fibre-optic cables.

Bagneux, part of the CNRS, France's national research agency. In addition, the device has to be cooled with liquid nitrogen. "I think the largest problem lies in achieving room-temperature operation," Gennser says. QCLs also emit too narrow a range of frequencies for use in long-distance optical telecommunications.

It remains unclear which approach will emerge as the best contender for lighting up silicon chips. But given the recent acceleration of progress, those intent on uniting the worlds of photonic and electronic information technology can afford to be optimistic. At the very least, silicon-based optoelectronics is starting to seem less of an oxymoron.

Philip Ball is a consultant editor of Nature.

1. Canham, L. T. *Appl. Phys. Lett.* **57**, 1046–1048 (1990).
2. Mizuno, H., Koyama, H. & Koshida, N. *Appl. Phys. Lett.* **69**, 3779–3781 (1996).
3. Hirschman, K. D., Tsybekov, L., Dutttagupta, S. P. & Fauchet, P. M. *Nature* **384**, 338–341 (1996).
4. Gelloz, B., Nakagawa, T. & Koshida, N. *Appl. Phys. Lett.* **73**, 2021–2023 (1998).
5. Ackakir, O. et al. *Appl. Phys. Lett.* **76**, 1857–1859 (2000).
6. Holmes, J. D., Johnston, K. P., Doty, R. C. & Korgel, B. A. *Science* **287**, 1471–1473 (2000).
7. Pavesi, L., Dal Negro, L., Mazzoleni, C., Franzò, G. & Priolo, F. *Nature* **408**, 440–444 (2000).
8. Dehlinger, G. et al. *Science* **290**, 2277–2280 (2000).
9. Curry, R. J., Gillin, W. P., Knights, A. P. & Gwilliam, R. *Appl. Phys. Lett.* **77**, 2271–2273 (2000).
10. Wang, T., Moll, N., Cho, K. & Joannopoulos, J. D. *Phys. Rev. Lett.* **82**, 3304–3307 (1999).
11. Zhang, P., Crespi, V. H., Chang, E., Louie, S. G. & Cohen, M. L. *Nature* **409**, 69–71 (2001).



Oxide formation: reaction details studied, reported in brief

Sir—Nineteen years ago, I published a sonnet in *Nature*¹ reporting new chemistry carried out at the University of Stirling. Although this application of the sonnet form was well received by critics, it has not subsequently been widely followed.

My present sojourn in Japan has caused me to consider the haiku. It is, with only 17 syllables in its English language form, of legendary brevity. The form is not extensively known outside Japan and the following is illustrative (although not original in content).

"War-time Deprivation and Hope"

*Yes! We have no bananas,
We have no bananas today.*

But soon?

On the positive side, this brevity provides little scope for waffle in the discussion section of a scientific publication but, on the down side, experimental details must be kept to a minimum (I rule out, of course, the option of stringing together haiku as not within the best traditions). The 17 syllables are normally 5, 7, 5 but some latitude is allowed, as with the usual allusions to nature, the seasons, and other conventions of form and content² (although purists might quibble).

Since the relatively recent discovery of the wide-ranging biological effects of nitric oxide (NO), considerable effort has been applied to the discovery of compounds which will liberate it under physiological conditions. Some *N*-nitrosohydroxylamines are such compounds³. However, we have discovered that closely related *N*-nitrosohydroxylamines undergo an alternative decomposition under very similar reaction conditions to liberate nitrous oxide, N₂O (refs 4,5). Moreover, this alternative reaction involves highly electrophilic intermediates analogous to ones involved in reactions of nitrosamines, for example, which are known to be seriously hazardous to human health. Investigation of the molecular basis of these alternative reactions of *N*-nitrosohydroxylamines was, therefore, of some importance.

As a further attempt to encourage the scientific community to explore alternative forms of scientific reporting, and to draw to the attention of a wider readership our recent work at Newcastle on the mechanisms of decomposition of *N*-nitrosohydroxylamines in which we identify the molecular basis of the alternative reaction channels referred to above⁶, I submit for consideration for publication in *Nature* the following.

"Decomposition of
N-nitrosohydroxylamines"⁴⁻⁶
Nitric or nitrous?
The leaving groups determine
Which oxide will form.

Howard Maskill

Institute for Fundamental Research in Organic Chemistry (IFOC), Kyushu University, Higashi-ku, Fukuoka, 812-8581, Japan

1. Maskill, H. *Nature* **294**, 606 (1981).
2. Henderson, H. G. *Haiku in English* (C. E. Tuttle, Rutland, 1967).
3. Saavedra, J. & Keefer, L. *Chem. Brit.* **36**(7), 30 (July 2000).
4. Maskill, H., Menneer, I. D. & Smith, D. I. *J. Chem. Soc., Chem. Commun.* 1855 (1995).
5. Haider, J., Hill, M. N. S., Menneer, I. D., Maskill, H. & Smith, J. *G. J. Chem. Soc., Chem. Commun.* 1571 (1997).
6. Bhat, J. I. et al. *J. Chem. Soc., Perkin Trans. 2*, 1435 (2000).

Varied ecosystems need different fire protection

Sir—Covington states in his Commentary¹ that the open ponderosa pine forests of the western United States are "in widespread collapse" because fire suppression by humans has eliminated the low-intensity surface fire regime that maintained the open, park-like structure of these forests. He fears this will lead to an "unprecedented" crown fire regime that will eliminate forests.

Although a crown fire regime in ponderosa pine forests may be unprecedented, there is a wide variety of ecosystems in North America that have always had a high-intensity crown fire regime, including chaparral in southern and central California, subalpine forests in the Rocky Mountains, and the boreal forest of Canada and Alaska², among others. Some forest managers and conservationists believe that wildfires in these crown fire regime ecosystems are more extreme in their behaviour and effects than ever before, but in most ecosystems there is little or no good evidence to support this belief³.

In the chaparral ecosystems of southern and central coastal California, for example, the crown fire regime has changed little since Europeans first settled the area, despite extensive fire suppression efforts over the past decades⁴. There has been an increase in the number of human-caused fires; these, however, have had an insignificant impact on the total area burnt each year. Consequently, high-intensity wildfires continue to dominate in these ecosystems and are not an artefact of fire management policy.

In boreal and subalpine forest ecosystems, it is often stated that, because of fire suppression, fire intensity will increase because fuel accumulates as forest stands age. However, while large fuels may accumulate, the fine fuels (1 cm diameter) that contribute most to fire intensity and spread reach an equilibrium well before

stand maturity. Thus, we should not expect a significant change in fire behaviour owing to "unnatural fuel accumulation" in these closed-canopied forests.

Although there may be evidence to suggest a shift from low-intensity surface fires to high-intensity crown fires in the ponderosa pine forests of the southwest, it would be prudent for resource managers from other western ecosystems to demand convincing evidence before applying Covington's suggestions to their areas.

Sheri L. Gutsell*, **Edward A. Johnson***,
Kiyoko Miyanishi†, **Jon E. Keeley‡**,
Matthew Dickinson*, **Simon R. J. Bridge§**

**Department of Biological Sciences, University of Calgary, Calgary, Canada*

†Department of Geography, University of Guelph, Guelph, Canada

‡US Geological Survey, Western Ecological Research Centre, Sequoia–King County National Park, USA

§Northeast Science and Information Section, Ontario Ministry of Natural Resources, South Porcupine, Canada

1. Covington, W. W. *Nature* **408**, 135–136 (2000).
2. Johnson, E. A., Miyanishi, K. & Bridge, S. R. J. *Conservation Biology* (in the press).
3. Anderson, J. E., Romme, W. H., Meyer, G., Knight, D. H. & Wallace, L. *Science* **283**, 175–176 (1999).
4. Keeley, J. E., Fotheringham, C. J. & Morais, M. *Science* **284**, 1829–1832 (1999).

Opposition to animal law is not lack of concern

Sir—A letter to the Editor from Bert van Zutphen (*Nature* **409**, 452; 2001) suggests that there is a difference between the views of the Federation of American Societies for Experimental Biology (FASEB) and the European Science Foundation on the issue of laboratory animal welfare. In fact, our positions are quite similar, as demonstrated by the principles for the care and use of animals in research and education adopted by the FASEB Board of Directors in 1994 (see <http://www.faseb.org/opar/animal1.html>).

Our opposition to proposed changes in the United States' Animal Welfare Act — extending it to cover rats, mice and birds — is not based on a lack of concern for animal welfare. In the United States, the care of most rats, mice and birds in medical research is subject to the Public Health Service policy on humane care or voluntary accreditation by the Association for the Assessment and Accreditation of Laboratory Animal Care, International.

What we object to is the additional level of bureaucracy that would divert resources from biomedical research without providing any new benefits for animals.

Mary J. C. Hendrix

Office of Public Affairs, FASEB, 9650 Rockville Pike, Bethesda, Maryland 20814, USA

In search of unity

Is the poetry in the science or in its appropriation?

Science and Poetry

by Mary Midgley

Routledge: 2000. 230 pp. £19.99

Philip Clayton

The thesis does not have to be new for the book to be interesting. One may already be convinced that science and poetry need not do battle, yet still read with fascination as Mary Midgley insightfully retraces some of the familiar lines. Of course, 'poetry' is itself a poetic way of putting her point, a metaphor that stands for all that humans value as beautiful, meaningful and significant. But *Science and Poetry* offers not merely poetic musings about the practice and conclusions of science; Midgley has bigger goals in mind: "This is a book about personal identity, about who and what we are. It is about the unity of our lives."

Midgley's thesis can be stated in a few simple sentences if one leaves out the poetry (although one should not rush to do so, as it may be the best part of the book); for example: "subjectivity is not scandalous." The major problem with the interpretation of science today, of which both practitioners and the general public are guilty, is that "the Cartesian split [between subject and object] is still rampant today". The world of our experience is composed neither of matter alone, as materialists hold, nor of matter and mind, each held in splendid isolation. Rather, the two exist in "symbiotic" relatedness, the exact balance between them established more by what is socially and personally useful than by any *a priori* distinction derived from either science or religion. The best — the highest, or most useful? — expression of this symbiosis is the so-called Gaia hypothesis proposed by James Lovelock — that the Earth is a vast, self-regulating organism. It can serve as a poetic inspiration, an ethical guideline, and a model for the interactions between humans and their world.

For some readers Midgley will be preaching to the choir; for others these claims will be controversial. The book is recommended reading for both groups: for the first because of the beauty of Midgley's prose and the innovative way she formulates the thesis; and for the second because the arguments are clearly and carefully formulated. Not all philosophers are equally helpful to science; this one merits our attention.

Midgley's case is credible because there has been a philosophical change of climate. This change has three fronts, which correspond to the book's three parts: a revised view of scientific rationality, an overwhelming rejection of mind-body dualism, and a more holistic understanding of the relationship

between humans and their environment. About the first little more needs to be said: logical positivism, by which the truth or falsity of a statement can only be assessed by empirical tests, has suffered a death as definitive as a philosophy can suffer. Human interpretations clearly play a role in the growth of science; dichotomies between 'purely objective' science and 'purely subjective' human existence have thereby become deeply suspect. To be a scientist you do not need to disdain poetry.

Midgley is on equally strong ground when it comes to mind and body. She shares the widespread suspicion about the old Cartesian dichotomy between them, yet resists the elimination of mind or its complete reduction to body. But her view also faces hard questions: do minds, or at least mental phenomena, really exist, or is a ticking off of the relevant physical mechanisms sufficient to make sense of what we call mental? Does nature consist of a whole series of emergent centres of activity which include, not only organisms, but species and ecosystems as well? Without an answer to the first question, one is unsure why science must retain, rather than explain, the level of the personal; and without an answer to the second, one is unsure what to make of the claim that the planet is a sort of quasi-organism (Gaia) above and beyond all the living things that reside on it.

Finally, a holistic approach to the relationship between humans and the ecosystem has indeed become the dominant background assumption in environmental studies today. 'Resource management' philosophies have become a symbol of all things evil — even if the scientific techniques of resource management continue under new names. Midgley knows that the environmental movement, both scientific and popular, now comfortably uses myths, stories and poetry from native peoples and the world's sacred scriptures to explain the interdependence of all organisms and to motivate responsible behaviour. In a sense, the use of such poetry represents the core strategy of her book, especially in the concluding chapters.

A certain Romantic idealism exudes from the pages, albeit often a charming one. At its weaker moments, though, one fears that the book should have been called *Sufficiency of Science or Romantic Holism?* rather than *Science and Poetry*. The prose comes closest to black versus white, and the arguments are weakest, when Midgley attacks claims to the sufficiency of scientific knowledge. "What [scientists] need to avoid," she



BRIDGEMAN ART LIBRARY

Apollo with one of his Muses.

writes, "is fundamentalism — the conviction that the particular imaginative vision espoused by their own party of current scientists is a solitary gospel which must always prevail." Inevitably, a Romantic poet's voice follows her stern warning, in this case, William Blake:

May God us keep

From Single vision and Newton's sleep!

But all is not effortless in Eden. It's not without reason that subjectivity and objectivity have battled their way throughout the modern period, and that science and poetry have inhabited sharply different realms.

Declaring their peaceful coexistence is not the same as extinguishing the slowly burning fuse. We need to know what *is* the place of the mental in the Universe. Perhaps Midgley would have benefited here from invoking the debates on the 'anthropic principle' — the cosmological principle that theories of the origin of the Universe are constrained by the need to allow individual human existence. Or greater attention to the more serious of the recent religion–science discussions might have helped. Then there's Gaia, a thesis that has understandably made scientists, and most especially biologists, wince. As an ethical metaphor it raises no problems; but is it true? Clearly, the planet as a whole lacks the structures of a cell or the organic interconnectedness of a higher organism. Lovelock's integration may be attractive: "For me, Gaia is a religious as well as a scientific concept, and in both spheres it is manageable. ... God and Gaia, theology and science, even physics and biology are not separate but a single way of thought." But there's some serious work to be done to make these claims stick.

Still, Midgley writes perceptively — and beautifully — about many things. She understands the power of mathematical physics without needing to deny that there are also "social facts". Just as philosophers need to get the world right, so scientists equally need to know the philosophical and ethical arguments that Midgley rehearses — even if matters are less black and white than she sometimes pretends. But, in the end, it is the poetry, including the poetry of Midgley's

prose, that makes the book worth reading. Like her, readers will sometimes have wondered with John Keats:

Do not all charms fly

At the mere touch of cold philosophy?

Yet they will agree with eighteenth-century poet and physician Mark Akenside that there is beauty and poetry in even the most rigorous understanding of nature:

Nor ever yet

The melting rainbow's vernal-tinctured hues

To me have shone so pleasing, as when first

The hand of science pointed out the path

In which the sun-beams gleaming from the west

Fall on the watery cloud. ■

Philip Clayton is in the Department of Philosophy, Sonoma State University, Rohnert Park, California 94928, USA.

Art in the plastic age

The Impact of Modern Paints

by Jo Crook & Tom Learner

Tate Gallery Publishing: 2000. 192 pp. £16.99

Mark D. Haw

The realization of ideas in art depends on the technology available to the artist. The twentieth century's major technological contribution was the invention of 'synthetic' paints. Water-based polymer

solutions and emulsions overcame the limitations of slow-drying, expensive oil paints. Although synthetic paints were originally developed for the household market, artists soon began to experiment with the new technology.

Synthetic paints are basically plastics in water. Thus, they dry by fast evaporation (oils dry by slower oxidation), and because the base solvent is water, a wider range of soluble pigments is available, improving the range of colours. The composition of the paint can be fine-tuned to provide properties such as drip resistance. And synthetic paints can be used on anything from paper to aluminium.

The Impact of Modern Paints investigates how artists have exploited the new paint technology. Jo Crook and Tom Learner, art conservationists at London's Tate Gallery, consider 10 artists from the Tate collection, including Andy Warhol, Bridget Riley and David Hockney. Between the 1950s and 1970s these artists were pioneers of the new paints. The practice of their art frequently veers towards science, their experiments in what could be done with the new synthetics being perhaps as important as their contributions to the conceptual development of art.

So the stage is set for a fascinating first-hand account of how Roy Lichtenstein, John Hoyland, Hockney and others have taken advantage of synthetic paint technology. But, disappointingly, the book contains few insights. Crook and Learner seem oddly unenthusiastic about their subject, and the text is spoilt by some particularly banal quotes from the artists. Obvious technical tricks are too often described in inordinate detail. And some comparison between synthetic and 'traditional' oil techniques would have been interesting, if the purpose really is to assess the impact of the new technology.

Nevertheless, the book does contain a brief but clear layman's account of the science of modern paint, and there are many excellent reproductions. Obliquely lit detail shots give a perspective unavailable to the casual gallery visitor, allowing an appreciation of texture as well as colour. The glossary is useful, if idiosyncratic ('paint' is defined, 'gesso' isn't).

The links between modern art and modern materials science are many and tantalizing. I suspect there is an intriguing story waiting to be told — but one that this book only hints at. Unfortunately, although they dab in a few interesting details, the authors have neglected to prepare the canvas, let alone apply the undercoat. ■

Mark D. Haw is in the Department of Physics & Astronomy, University of Edinburgh, James Clerk Maxwell Building, Mayfield Road, Edinburgh EH9 3JZ, UK.



David Hockney's *A Bigger Splash*: the artist took advantage of new paint technology.

From the monastery to the laboratory

In Mendel's Footnotes: An Introduction to the Science and Technologies of Genes and Genetics from the 19th Century to the 22nd

by Colin Tudge
Jonathan Cape: 2000. 354 pp. £18.99

Brian Charlesworth

Genetics has made the most spectacular progress of any part of biology over the past century, developing from an obscure science practised by a handful of people to one that is referred to almost daily by the media. The concepts and techniques of genetics permeate all fields of biology; the potential benefits and threats of its applications to agriculture and medicine are the objects of often-strident debate. And it is the only branch of science whose practitioners have been the subject of intensive political persecution, as a result of Trofim Lysenko's malign influence in the Soviet Union. For these reasons, it is vitally important that our knowledge of genetics is communicated to the general public. However, those who have struggled to teach it to undergraduates may view such an enterprise with some scepticism.

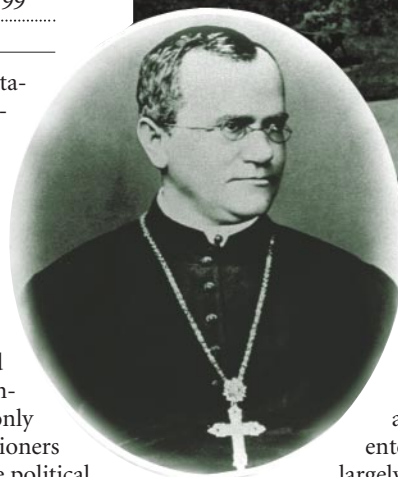
Colin Tudge's *In Mendel's Footnotes* is the latest in a series of popular books on genetics. It adopts a quasi-historical approach in order to emphasize how Gregor Mendel's ideas shaped the development of modern genetics. Tudge devotes a good deal of space to Mendel and his influence, with two chapters devoted largely to biographical details of the man and his family, and the intellectual environment to which he was exposed.

Mendel's story is, of course, one of the most romantic in the history of science. A Catholic monk in a provincial town of the Austro-Hungarian empire made the crucial breakthrough in our understanding of inheritance in terms of the transmission of discrete hereditary factors, and was ignored for 35 years. Of course, others (notably Francis Galton and August Weismann) came close to deducing the principles of mendelian inheritance, and others (including Charles Darwin) reported mendelian ratios in experimental crosses. As R. A. Fisher pointed out in his famous 1936 analysis of Mendel's 1866 paper, it is virtually certain that Mendel formulated his theory of particulate inheritance before designing his meticulous experiments; this was crucial to his success where others had failed.

Tudge's account of Mendel's background



Fertile soil: Mendel's monastery garden at Brno, where he performed the experiments that laid the foundations of modern genetics.



as a poor boy who entered the Church largely for economic reasons is entertaining and instructive, even if the importance of fruit-growing in nineteenth-century Moravia is perhaps overemphasized. Tudge also gives a clear account of Mendel's experiments and their implications, although the absence of diagrams probably makes it hard for the general reader to follow.

Tudge's treatment of post-mendelian genetics has serious weaknesses. They flow from his stated wish "to show that all the modern ideas of twentieth-century classical genetics flow naturally from Mendel's initial observations". This is surely going too far. The geneticist's single most important tool is genetic mapping, which depends on the post-mendelian concept of linkage. This is treated superficially in a couple of paragraphs. And there is no account of how genes can be ordered genetically and physically. The reader is therefore left with no idea of how we know that genes are located in a linear array on chromosomes. Although restriction enzymes are mentioned, the reader is not told how genes with known effects but unknown products can be isolated and cloned. This is, of course, crucial to the characterization of most genes that have important roles in development, including those responsible for major human genetic disorders.

Tudge rightly draws attention to the role of mendelian genetics in underpinning our understanding of evolutionary processes, and the revitalization of evolutionary biology that resulted from the application of genetics. But he focuses almost entirely on natural selection in the context of the 'selfish

gene' approach. There is no reference to the fact that random fluctuations can occur in the frequencies of variants in finite populations, surely the most important single advance on Darwin that has been made in our understanding of the mechanism of evolution (neither Motoo Kimura nor Sewall Wright is mentioned). Given the central role that this is likely to have played in the evolution of DNA sequences, and current interest in the use of sequence data for reconstructing human evolution, this seems an extraordinary omission.

Tudge also goes overboard on 'evolutionary psychology' as a basis for interpreting human behaviour in terms of natural selection. And he ignores the difficulties of testing biological explanations of behaviour in a human context. It is disconcerting to see hypotheses such as the selective advantage of infanticide by stepfathers presented on equal terms to the exceptionally rigorously tested concepts of classical and molecular genetics. I feel this conveys the wrong impression to the non-scientist, who often finds it difficult to distinguish between what has been firmly established and what is the subject of debate and ongoing research.

The final part of the book is devoted to animal and plant improvement, biotechnology and the ethics of genetic manipulation. This is mostly sensible stuff, making it clear that it is a matter for informed choice as to how far and how fast to proceed. It should be a useful guide for anyone concerned with these issues. Such a reader will probably not be troubled by the omissions I have noted, nor by the large number of minor factual inaccuracies in the book — the incidence of haemophilia is not increased by inbreeding, as it is caused by a sex-linked recessive gene;

book reviews

not all Hox genes are closely linked; François Jacob and Jacques Monod discovered the operon in *Escherichia coli*, not *Lactobacillus*; and so on. Popularizers of science have an obligation to get their facts right, and the quality of this book is impaired by the failure to do so. ■

Brian Charlesworth is at the Institute for Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, UK.

Corruption and its consequences

Captive State: The Corporate Takeover of Britain

by George Monbiot

Macmillan: 2000. 430 pp. £12.99

Frances Cairncross

Whatever you think of George Monbiot's political views, he writes like an angel. Whether on the Isle of Skye, observing Robbie the Pict "with a broad ginger beard and hands like shovels", or walking along the River Usk, where "an old fisherman, gaunt as a heron, stalked up the bank", his book is full of poetic little pictures. Most of them describe the ordinary people who are the victims of the collusion between government and business that he so deplores.

As one of Britain's best campaigning authors and environmentalists, Monbiot has used his powerful pen mainly against the corruption of foreign governments in such sinister parts of the world as the Amazon, Kenya and Indonesia. Now he is describing what he sees as corruption in Britain: the insidious links between government and business which, he believes, is corrupting the political process, capturing politicians and undermining democracy. Corporations, he seethes, "are seizing powers previously invested in government and using them to distort public life to suit their own ends".

In a powerful paragraph, he announces: "The struggle between people and corporations will be the defining battle of the twenty-first century. If the corporations win, liberal democracy will come to an end. The great social democratic institutions ... will be toppled." Scary stuff, and very much in the mood of the 1999 demonstrations against the World Trade Organization in Seattle.

Yet many of the examples in Monbiot's book are testimony, not to the wickedness of big companies, but to the abysmal ineptitude of bureaucrats and the cynicism of politicians. Take his extremely readable chapter on the scandal of the Skye Bridge. The bridge, which links mainland Scotland with the Isle of Skye, was built largely with private money in exchange for the right to levy a ludicrously high toll, not just on tourists, but also on the



Whenever government touches companies, there is an opportunity for pressure — and corruption.

unfortunate inhabitants of the island.

The villains here are not the Bank of America, which apparently owns the company (although it could do with some public-relations advice here), but the public servants who signed the initial deal and the politicians who, having savaged it while in opposition, then lamely backed down when in power.

Sometimes, the villains are even closer to home. One of Monbiot's pet hates is the out-of-town superstore. He chronicles the way supermarket companies extract planning permission by bribing local councils with promises of building new homes or sports centres. Monbiot is eloquent about the misery that big stores inflict on farmers and small shops. Yet is this really about corruption? Surely, the real problem is that people desert their local shops and drive out of town because of the range of products and low prices the superstores can offer. If they were willing to pay a little more and accept less choice, it would be the superstores, not the local butcher and baker, that would go out of business.

So is there no case to answer? On the contrary; wherever government touches companies — where it can grant or withhold permission, pay a subsidy, allocate a contract — there is an opportunity for pressure, profit — and corruption. Each transaction puts temptation in somebody's way. The solution, clearly, is to reduce such contacts. But no politician thinks quite that way. After all, the opportunities to grant or withhold favours are an aspect of power and privilege that makes politics worthwhile. But two such contacts feature so often in Monbiot's pages that they surely call for drastic reform.

One is the UK Private Finance Initiative. Not only is this at the heart of the Skye Bridge debacle; it leads to bad decisions on

the building of houses, hospitals and roads. In essence, it is an elaborate way for the public sector to get around its own borrowing controls, by getting private companies to do the financing. But, as with all subterfuges, there is a price to be paid, in terms of cost and complexity. Monbiot says the Private Finance Initiative should be scrapped. He is right.

The second corrupting contact is the granting of planning permission. Monbiot's pages are full of stories of large companies that sway local politicians to override their voters' protests. He wants to allow objectors the right to appeal against developers, as developers can appeal against successful objectors. He's right again — although a better reform might be to simplify the process — by setting out a few universal rules on height, density and so on — that could not be overridden for particular projects. The more open a process is to political intervention, the greater the scope for abuse.

For most senior managers of large companies, running a business is not about trying to subvert democracy. Many chief executives would argue that government is not too weak but too powerful — too able to intervene and arbitrarily change the rules. Indeed, the pressures on companies to do deals with government rarely come from a desire to topple democracy or infiltrate government. Most large public companies simply want to stay in business, make money and, above all, sell things to their customers. Their customers — which frequently means you and me — are the true power behind the evil empire. If we don't buy, companies don't sell. In a democracy, the vote is a powerful weapon, but the power of the purse is greater still. ■

Frances Cairncross is at The Economist, 25 St James Street, London SW1A 1HG, UK.

Lombroso and Tolstoy

An anthropologist's unwitting gift to literature.

Paolo Mazzarello

More than 7,000 participants attended the twelfth International Medical Congress of Moscow in August 1897. Among the prominent scientists there was Italian anthropologist Cesare Lombroso, who presided over a session dedicated to mental illness.

Lombroso was world famous for his theory that genius was closely linked with madness. According to him, genius and madness were two faces of the same psychobiological reality — as in a non-Euclidean space in which the two extremes touch. A man of genius was a degenerate, an example of retrograde evolution, in whom madness was a form of biological compensation for excessive intellectual development.

This regression, the theory stated, produced its own phenotypic stigmata, such as the cranial asymmetry of Pericles, Kant and Dante; the sub-microcephaly of Descartes; the small stature of Horace, Plato and Epicurus. According to Lombroso, the thirteenth-century scholar and patron saint of natural scientists, Albert the Great, “was of such small stature that as he kissed the Pope’s foot, the Pontiff ordered him to stand up, thinking he was on his knees”¹.

Lombroso’s participation in the Moscow Congress inspired him to test his theory about the pathology of genius. Why not meet Leo Tolstoy, the supreme genius of world literature, in his natural habitat, to scrutinize his features and confirm his theories by seeing Tolstoy’s degenerative aspects with his own eyes? Like a naturalist embarking on a tour of South America to test his evolutionary theories, Lombroso decided to take himself to Tolstoy’s home, known as Yasnaya Polyana. For Lombroso the writer represented the “true disguised genius of alienation ... in whom, one might say, the sicker the body, the more sublime the [intellectual] products”. He imagined Tolstoy would be “cretinous and degenerate-looking”¹.

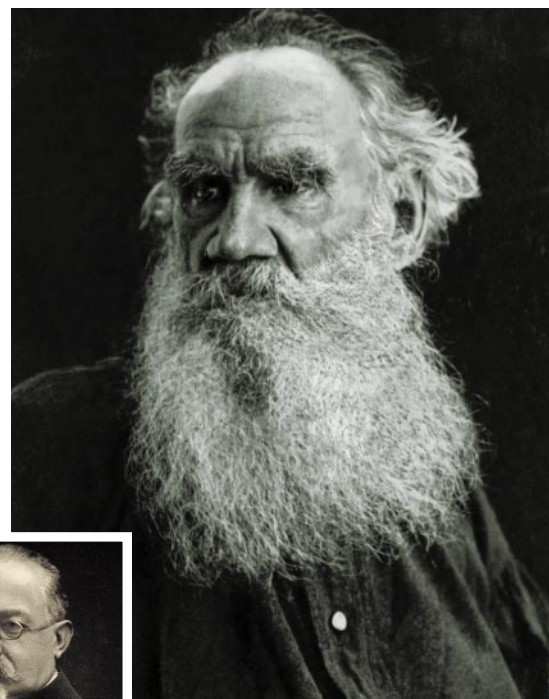
On arriving at Tolstoy’s house, Lombroso found himself face to face with a soldierly-

looking old man, whose penetrating eyes and severe bony face seemed more like those of a good, solid peasant who had served in the army than those of a thinker. Physically there was nothing degenerate about him. His attitude was calm, correct and friendly; there was certainly method in his ‘madness’. But when the conversation took a turn that opposed Tolstoy’s ideas, Lombroso realized the “total impossibility of engaging him in debate without irritating him”². In particular, they argued about the criminological issue closest to Lombroso’s heart: the theory of the born delinquent. According to this theory, certain types of criminals had suffered arrested development at an early stage and were therefore the most ‘atavistic’ type of human being, without hope of reform. Society has the right to defend itself from this kind of delinquent, argued Lombroso, even by means of the death penalty, just as a man defends himself from wild animals without blaming them for not having been born lambs.

Tolstoy remained deaf to these arguments. “He knitted his terrible eyebrows,” Lombroso records, and hurled menacing glances from his deep, penetrating eyes. Finally he erupted: “All this is nonsense! All punishment is criminal!” Human beings have no right to judge their fellows, Tolstoy continued, and no form of violence is admissible under any circumstances.

Lombroso could not understand this attitude, which he found eccentric: the product of a sick and violently passionate mind. Whereas on the evening of 27 August 1897, Tolstoy noted in his diary: “Lombroso came. He is an ingenuous and limited old man.” (At 61, Lombroso was seven years younger than Tolstoy.) In January 1900 he added, about Lombroso’s theories: “All this is an absolute misery of thought, of concept and of sensibility.”

Some months after this visit, Tolstoy rewrote his last great novel, *Resurrection*, which had previously only existed in draft form. In the definitive text of *Resurrection*, Tolstoy added, among other things, a detailed description of the legal process and punishments current in Russia at the end of the nineteenth century, in which Lombroso’s anthropological theories were discussed and rejected as immoral. Delinquency was not “evidence of degeneration of a delinquent



Age of reason: Lombroso (left) expected Tolstoy (above) to be “cretinous and degenerate-looking” because of his genius. Tolstoy found Lombroso “an ingenuous and limited old man”.

CORBIS

type of monstrosity, as certain obtuse scientists explained them to the government’s advantage”, he wrote. When prince Dmitry Nekhlyudov (alias Tolstoy), the main character, seeks an answer to the problem of criminal deviance in Lombroso’s books, he finds that the more he reads and the more attentive he is to Lombroso’s words, the more disappointed he becomes. During a trial the public prosecutor quotes Lombroso and the latest ‘scientific’ theories on heredity, evolution and the born delinquent, to support his case against a prostitute falsely charged with murder. “He’s going too far,” the president of the court comments to a colleague. “A very stupid fellow,” the colleague agrees.

Lombroso’s visit to Tolstoy was a failure on the intellectual plane. The genius and the madness of the two men had not been able to come into contact because they were too divergent to find a point of mutual understanding. However, this ‘scientific’ episode produced a lasting literary mark on *Resurrection* — the last great Tolstoy novel. ■

Paolo Mazzarello is at the Volta College, University of Pavia, and at the IGBE-CNR, Via Abbiategrosso 207, 27100 Pavia, Italy. His account of Lombroso and Tolstoy, *Il genio e l’alienista* (Bibliopolis, Naples, 1998), is soon to be translated into English by David Mendel.

1. Lombroso, C. *L'uomo di genio* (Bocca, Torino, 1894).
2. Lombroso, C. in *Tolstoj nelle memorie dei contemporanei* (ed. Opul'skaya, L.) 185–189 (Raduga, Moscow, 1984).

This episode produced a lasting literary mark on *Resurrection* — the last great Tolstoy novel.

The artistry of nature

Eshel Ben-Jacob and Herbert Levine

The endless array of patterns and shapes in nature has long been a source of joy and wonder to laymen and scientists alike. Discovering how such patterns emerge spontaneously from an orderless and homogeneous environment has been a challenge to researchers in the natural sciences throughout the ages. In 1610, the astronomer Johannes Kepler was already captivated by the beautiful shapes of snowflakes, perhaps the most striking examples of pattern in inorganic azoic (non-living) systems. But the origins of their six-fold beauty eluded him — Kepler lived too early to know about atoms and molecules. He did have the insight, though, that the symmetry of snowflakes resulted from an inherent power in matter, which he dubbed the “*facultas formatrix*”. Kepler was not alone in his inability to explain those graceful forms.

Only during the past two decades have the principles of transfer of the microscopic, molecular information to the macroscopic level of the visible flakes been deciphered. In

this case, we now understand how nature chooses one pattern over the other. But what about other pattern-forming systems? Indeed, many diverse out-of-equilibrium processes result in the emergence of patterns, such as spiral waves in the Belousov–Zhabotinsky redox reaction, Liesegang rings in reaction–diffusion systems, Rayleigh–Benard convection cells in heated fluids and disordered-branching patterns during electrochemical deposition. We believe that underlying these disparate patterns there is a set of overarching principles that can lend a unified perspective to this field of study. Recent progress towards such a perspective hints at the possibility of obtaining radically new insights into the even harder problem of pattern formation in living systems.

Patterning via competition

Pure substances at equilibrium (closed systems) usually assume a homogeneous (patternless) state or, at most, a simple periodic one. Back in the early 1950s, Alan Turing understood that complex patterns would emerge only in a system driven out of equilibrium (open systems), where there

Pattern

“Underlying many disparate patterns, we believe there is a set of overarching principles that can lend a unified perspective.”

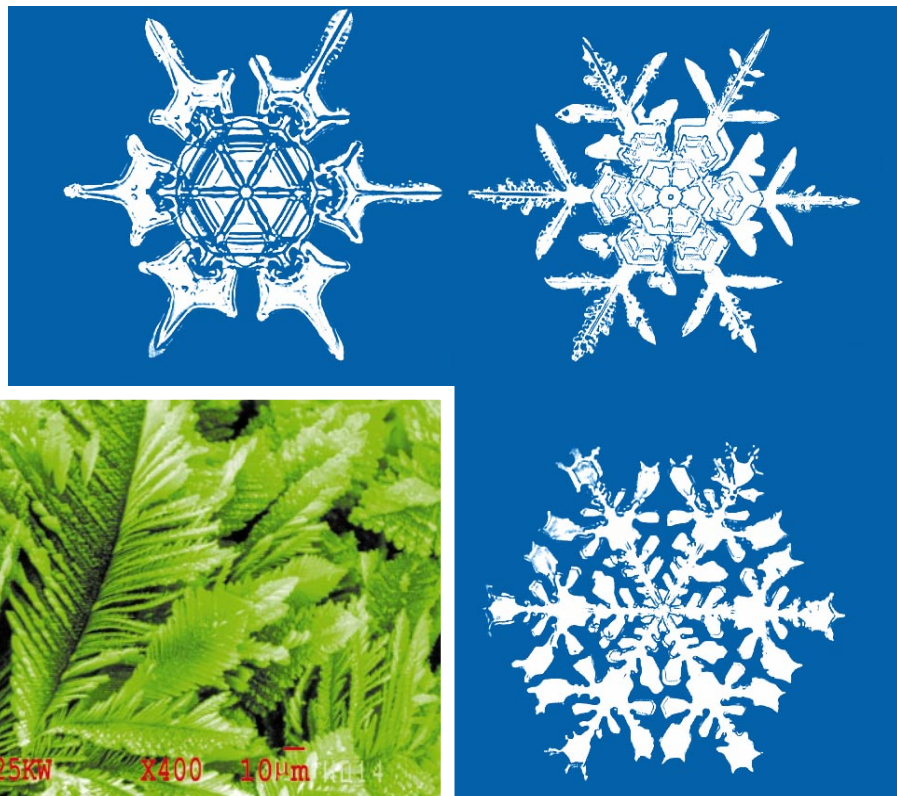
exists competition between various tendencies. For example, in snowflake formation the competition is between the diffusion of water molecules towards the flake and the microscopic dynamics of crystal growth at the solid–vapour interface. The diffusion kinetics tend to drive the system towards decorated and irregular shapes, with maximal interfacial area. The microscopic dynamics, giving rise to surface tension, surface kinetics and growth anisotropy, compete with this tendency and thereby impose characteristic length scales and overall symmetries on the resultant patterns. In other examples, competition between short-range activation and long-range inhibition, or between macroscopic heat transfer versus short-range viscous dissipation, has a corresponding role.

Micro–macro singular interplay

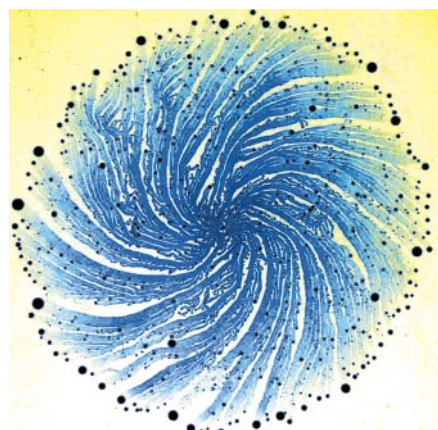
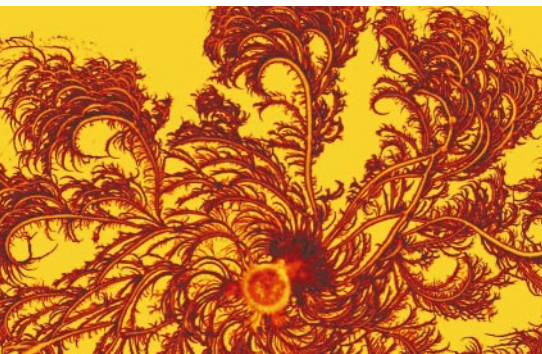
The aforementioned competition often gives rise to a two-way transfer of information between the microscopic and macroscopic scales. This is most obvious in the snowflake, in which the six-fold symmetry of the underlying lattice is manifest in the dendritic branches of the flake on the macroscopic (observed) level. At present, we understand that whenever the microscopic dynamics act as a singular perturbation (stabilizing competitor), details at the microstructural scale, such as preferred growth directions, will be amplified in effect by the macroscopic process. Chirality, the difference between left- and right-handed shapes, can act in a similar manner. By the same token, the macroscopic dynamics can reach down and affect the microstructure; changing the macroscopic conditions can force the small-scale structure to change, by favouring a particular growth mode over other possibilities. In other words, the macro-level and the micro-level organization must be determined in a self-consistent manner.

Morphology diagrams

The micro–macro interplay varies as the control parameters are changed. Because of the large degree of cooperativity in the pattern-forming process, we expect in general that there will be sharp transitions between different ‘essential’ shapes. Each of these shapes, or morphologies, represents a differ-



Patterns in non-living systems. Photos of ‘captured’ real snowflakes, taken by Wilson A. Bentley (Jericho Historical Society). Note the high level of (six-fold) symmetry together with the complexity of the patterns. Inset, ‘metal leaves’ produced during the electrochemical deposition of ZnSO_4 ; picture taken using an electron microscope (magnification $\times 400$). Both pictures show dendritic patterns and demonstrate that similar patterns can be seen in different systems and over very different length scales.



Complex patterns exhibited during colonial cooperative self-organization of *Paenibacillus dendritiformis* (top) and *Paenibacillus vortex* (bottom) show chiral asymmetry (all the branches have a twist with the same handedness). This observed chirality on the macroscopic (colonial) level results (via singular interplay) from the chirality of the flagella of the bacteria (the micro level). *P. vortex* shows organization of vortices (dots) composed of thousands of bacteria, which all circulate around a common centre. The delicate balance between order and chaos persists over many length scales, lending an unmistakable aesthetic quality to these images.

ent balance between the various competing tendencies leading to the formation of the pattern. Lending support to this notion is the well-studied example of diffusion-controlled growth. The same morphologies appear repeatedly in different systems exhibiting this underlying pattern-forming competition, with length scales ranging from micrometres to metres. This perspective brings to mind the idea of a morphology diagram, by analogy with a phase diagram for systems in equilibrium. In equilibrium, for given conditions, the phase that minimizes the free energy is selected and observed. The existence of an equivalent principle for dynamic non-equilibrium (open) systems is the most profound unsolved question in the study of pattern formation.

The power of cooperation

Among non-equilibrium systems, living organisms are the most challenging ones scientists can study. Although pattern for-

mation exists throughout the biological world, cooperative microbial behaviour seems a natural choice of a starting point to apply the lessons learned from azoic systems to living ones. Bacteria are the simplest organisms, yet a wealth of beautiful patterns are formed during colonial development of various bacterial strains. Some of the observed spatio-temporal patterns are reminiscent of those observed in non-living systems. Others exhibit an even richer behaviour, reflecting the additional layers of complexity involved in colonial development.

As in non-living systems, patterns emerge from the singular interplay between the micro level (the individual cell) and the macro level (the colony). That is, there must be an internal consistency between the microscopic interactions brought about by single-cell behaviour and the overall macroscopic organization of the colony as a whole. The building blocks of the colonies are themselves living systems, each having its own autonomous self-interest and internal degrees of freedom. At the same time, efficient adaptation of the colony to adverse growth conditions requires self-organization on higher levels — function follows form — and this can be achieved only via cooperative behaviour by the individual cells.

Thus, bacteria have developed sophisticated cooperative behaviour and intricate communication capabilities, including: direct cell-to-cell physical interactions via membrane-bound polymers, the collective production of extracellular 'wetting' fluid for movement on hard surfaces; long-range chemical signalling, such as quorum sensing; and chemotactic signalling, the collective activation and deactivation of genes, and even the exchange of genetic material.

The communication capabilities enable each bacterial cell to be both an actor and a spectator (using Niels Bohr's expression) during the complex patterning. The single-cell dynamics determine the macroscopic pattern even as it itself is shaped by that self-same pattern. For researchers in the pattern-formation field, the communication, regulation and control mechanisms that ultimately control the observable morphologies offer a modelling challenge that far surpasses that considered to date within the context of non-living processes. It should be evident to microbiologists that colonies have sophisticated capabilities for coping with hostile environmental conditions, capabilities that cannot be studied by focusing exclusively on the behaviour of the single bacterium.

Clues about complexity

Understanding pattern formation is intimately related to understanding the notion of complexity in open systems. Complexity is an oft-used word that still lacks any precise definition. Structural complexity might

refer to patterns with repeating yet variable units; in this sense, completely disordered structures are as simple as perfectly repeating ones. Functional complexity might be related to systems whose dynamic properties are not simply explained by detailed understanding of the constituent parts, perhaps because of feedback from the macro level. Unfortunately, neither of these intuitive notions has led to an objective operational measure.

An essential question in complex systems is the extent to which one can formulate theories that permit sensible predictions of the macroscopic behaviour of open systems without having to simulate in mind-numbing detail all the microscopic degrees of freedom. In physical systems in equilibrium, we are typically confronted with this question in the context of a two-level micro-macro interplay. We deal with this via the introduction of the entropy as an additional variable on the macro level. The entropy is a measure of (the logarithm of) the number of possible microscopic states for a given macro state of the system. Hence, it can be viewed as either our lack of information about the micro-level (looking from the macro level) or as the freedom in the microdynamics for given imposed macroscopic conditions (looking from the micro level).

Might complexity, properly defined, replace entropy as a fundamental property of macroscopic open systems? It is certainly intriguing that such systems tend to evolve in the direction of increased complexity as they are driven further from equilibrium. Future work on patterns, especially in living organisms, will no doubt offer some needed clues.

Eshel Ben-Jacob is in the School of Physics and Astronomy, Tel Aviv University, 69978 Tel Aviv, Israel.

Herbert Levine is in the Department of Physics, University of California, San Diego, La Jolla, California 92093, USA.

FURTHER READING

Ball, P. *The Self-made Tapestry: Pattern Formation in Nature* (Oxford Univ. Press, 1999).

Kessler, D. A., Koplik, J. & Levine, H. Pattern selection in fingered-growth phenomena. *Adv. Phys.* **37**, 255 (1988).

Ben-Jacob, E. & Garik, P. The formation of patterns in non-equilibrium growth. *Nature* **343**, 523–530 (1990).

Ben-Jacob, E. & Levine, H. The artistry of microorganisms. *Sci. Am.* **279**(4), 82–87 (1998).

Ben-Jacob, E., Cohen, I. & Levine, H. The cooperative self-organization of microorganisms. *Adv. Phys.* **49**, 395–554 (2000).

Correction: In the Millennium Essay "A cellular cornucopia" (*Nature* **408**, 773; 2000), the lizard was mistakenly cited in place of the newt in the context of limb regeneration.

The bite of *Allosaurus*

Gregory M. Erickson

A combination of analytical methods used in engineering and scanning techniques from medicine can be applied in some surprising areas of research — the biomechanics of dinosaur feeding, for instance.

Understanding the biomechanics of organs in the body requires a diversity of information¹. You have to know the structure's material properties (its stiffness, for instance), its size and shape, and its mechanical relationships with other structures. Data on the forces applied during use are also essential. It is possible, but by no means easy, to achieve this intricate level of analysis with living organisms. Think, then, how difficult the task is with a 150-million-year-old dinosaur, which the vagaries of time have left only as a skeleton.

As they show on page 1033 of this issue, however, Rayfield *et al.*² have achieved just such a feat in their biomechanical analysis of feeding in the Jurassic theropod *Allosaurus fragilis*. This was a carnivorous bipedal cousin of *Tyrannosaurus rex*, armed with 80 or more sabre-like teeth. Little is known of the natural history or physical capacities of this animal. Tooth-mark evidence shows that giant sauropods such as *Apatosaurus* were occasional meals³, and numerous injuries attest to the species' rough-and-tumble lifestyle⁴.

Rayfield and colleagues have created dramatic images of the *Allosaurus* skull (Fig. 1), from which they surmise a great deal about the functional morphology and feeding style of this little-understood animal. Perhaps more importantly, the results point the way towards a comparative understanding of biomechanical evolution in dinosaurs. Application of these techniques to a variety of species could reveal the biomechanical significance of the many cranial and dental forms found in dinosaurs and the ecological factors, such as changes in food types and availability, that fuelled the evolution of this diversity in form.

Until recently the only ways of characterizing the internal anatomy of dinosaur bones were by sectioning the bones or examining their broken ends. Furthermore, extracting a skeleton from the surrounding rock matrix required physically removing it. Because of the likelihood of damaging the specimen, this is often not attempted with intact skulls in which the matrix has filled the numerous cavities. In the past few years, palaeontologists have turned to medical and industrial computerized tomography (CT) as a non-destructive way of distinguishing bones and teeth from rock, and studying the inner and outer morphology of fossils⁵ (Fig. 2).

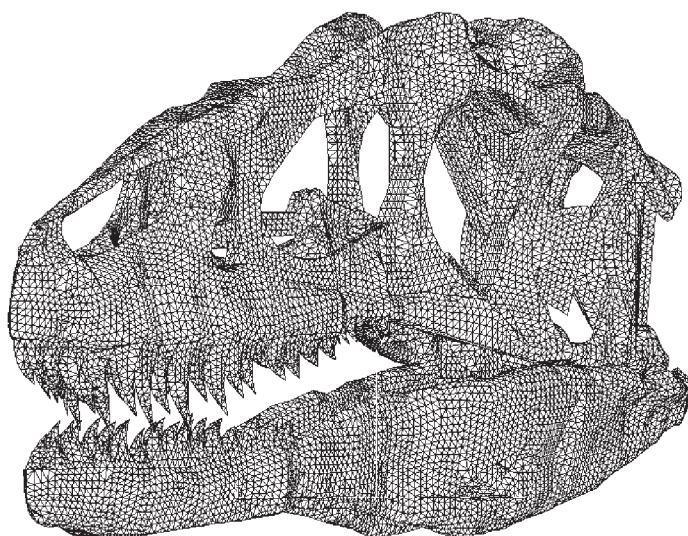


Figure 1 The *Allosaurus* skull, modelled by Rayfield *et al.*². The image is created from the conversion of computerized tomography data into a finite-element mesh.

Computerized tomography involves taking a series of X-rays to measure density differences in objects, allowing two- or three-dimensional internal profiles to be made. So it is possible to produce three-dimensional computerized images of fossil bones, and even entire skeletons.

Rayfield *et al.* have taken the approach one step further. Using advanced data-analysis techniques, such as those applied in

orthopaedic biomechanics⁶, they converted CT data for an entire *Allosaurus* skull into a 'finite-element mesh'. This mesh, which amounts to a mathematical description of the skeletal geometry, served as a foundation for computational mechanical analysis. When combined with information on the material properties of bone, and the forces theoretically generated by the jaw muscles, this approach allowed assessment of how

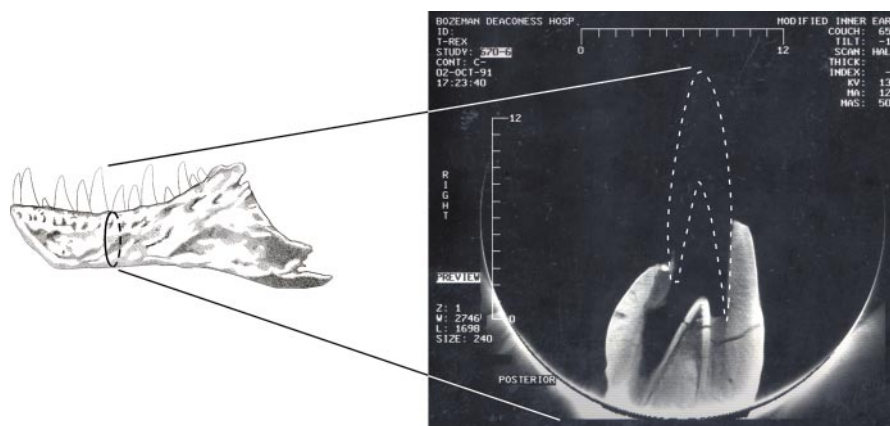


Figure 2 Computerized tomography in practice. In this CT scan¹² of a theropod dinosaur jaw, the osseous tissue (white) can be distinguished from the rock matrix (black). From scans like this it is possible to reconstruct the inner and outer geometry of fossil bones in three dimensions without destructive analysis. This particular scan reveals the location and degree of development of a replacement tooth (the conical white structure seen deep within the jaw).

forces and stresses were transmitted throughout the *Allosaurus* skull during feeding. The overall result is an impressive biomechanical model of the skull's form, function and performance.

The model is extraordinary in its three-dimensional accuracy. Furthermore, it is made at a regional (skull and jaw) anatomical level, rather than a more common, single-bone level. Detailed skull-and-jaw finite-element analysis such as this has been attempted only rarely for living organisms, for instance with the human models used in crash tests by the automobile industry. To date, the few applications in vertebrate palaeontology have been limited to descriptions of single-bone mechanical loading in which stress distributions, not stress magnitude, were studied^{7,8}.

Rayfield and colleagues' analysis indicates that the maximal bite forces of the 1.4-tonne *Allosaurus* were surprisingly low, on par with those for much smaller living carnivorous mammals such as wolves and leopards⁹, and at least six-times lower than sub-maximal estimates for its larger cousin, *Tyrannosaurus rex*¹⁰. This by no means suggests that the beast did not pack a lethal punch, only that it was discriminating in where and how it bit its prey. Rayfield *et al.* contend that *Allosaurus* had a feeding strategy reminiscent of that of Komodo dragons, inflicting rapid, slashing bites primarily to the soft tissues which meant that the prey bled to death. *Tyrannosaurus rex*, on the other hand, with its much more robust teeth, was capable of smashing through flesh and bone, and delivering deeper and perhaps more instantly lethal bites¹⁰.

Contrary to the general view that the *Allosaurus* skull was delicate, the authors argue that it was extraordinarily strong relative to the bite force — that is, it was 'overbuilt'. The strut-like skull structure was mechanically well suited for dissipating biting forces and stresses. Rayfield *et al.* propose that the overbuilt skull was adapted for absorbing forces from impacts with prey, and from drawing its teeth through that prey. For the moment, however, it cannot be said whether this overbuilt skull structure reflects these or other mechanical effects, non-mechanical influences (such as accommodation of soft tissues or sexual display), or simply the typical reptilian condition.

Rayfield and colleagues' approach can be used to tackle other questions in dinosaur biology. Were overbuilt skulls common to all theropod dinosaurs? Which biomechanical or other factors led to the development of this and different skull morphologies? What was the mechanical significance of the intracranial joints in dinosaur skulls? Further loading tests with this model, as well as bite-force simulations using full-scale models of dinosaur teeth and jaws, might be used to address some of these questions.

More generally, the CT-scan/finite-element modelling method could be applied to within- and between-species analyses. Topics might include changes in physical capacities during animal development and evolution; the effects of morphology and scale on biomechanics; and the functional significance of anatomical form in extinct animals (many of which cannot be modelled by living analogues). These exciting new areas of research require expertise from such disciplines as mechanical engineering, vertebrate palaeontology, evolutionary biology and comparative morphology. The trend towards an integrative approach in many biology departments over the past decade has encouraged a multidisciplinary approach¹¹. Advances in computer technology, and the emergence of a new breed of scientist versed in these various fields, point to a rosy future for biomechanical analyses of extinct organisms. ■

Structural biology

A 3D view of sodium channels

William A. Catterall

The voltage-gated sodium channel is essential for nerve cells to function. But how does it work? A low-resolution, three-dimensional structure provides some provocative insights.

Sensation, emotions, thought and movement all depend on voltage-gated sodium channels — transmembrane proteins that initiate action potentials in nerve, muscle and other electrically excitable cells. The hallmarks of sodium channels include rapid activation and inactivation when the voltage across the plasma membrane of an excitable cell is depolarized (voltage-dependent gating), and efficient and selective conduction of sodium ions through them. On page 1047 of this issue¹, Sato and colleagues provide the first three-dimensional view of the sodium-channel protein. The images reveal unexpected structural features that may contribute to the protein's function in electrical excitability.

The α -subunit of the sodium channel, which forms the ion-conducting pore, was initially identified as an unusually large polypeptide (relative molecular mass, 260,000) by labelling with specific neurotoxins^{2,3} that modify the generation of action potentials. Subsequent DNA-cloning experiments revealed the amino-acid sequences of a family of nine related α -subunits in mammals^{4,5}. Each α -subunit contains four homologous domains, I to IV, each with six predicted transmembrane segments (Fig. 1).

Further work provided a two-dimensional molecular map that showed which structural features are responsible for which

Gregory M. Erickson is in the Department of Biological Science, Florida State University, Conradi Building, Tallahassee, Florida 32306, USA. e-mail: gerickson@bio.fsu.edu

1. Beaupré, G. S. & Carter, D. R. in *Biomechanics: Structures and Systems* (ed. Biewener, A. A.) 149–174 (IRL, Oxford, 1992).
2. Rayfield, E. J. *et al.* *Nature* **409**, 1033–1037 (2001).
3. Hildebrand, S. F., Gilmore, C. W. & Cochran, D. M. *Cold-Blooded Vertebrates* (Smithsonian Science Series, New York, 1930).
4. Laws, R. R. *J. Vert. Paleont.* **17**, 59A–60A (1997).
5. Chapman, R. E. in *The Complete Dinosaur* (eds. Farlow, J. O. & Brett-Surman, M. K.) 112–135 (Indiana Univ. Press, Bloomington, 1997).
6. Keyak, J. H., Meagher, J. M., Skinner, H. B. & Mote, C. D. *Jr J. Biomed. Eng.* **12**, 389–397 (1990).
7. Rensberger, J. M. in *Functional Morphology in Vertebrate Paleontology* (ed. Thomason, J. J.) 151–172 (Cambridge Univ. Press, 1995).
8. Carter, D. R., Mikic, M. & Padian, K. *Zool. J. Linn. Soc.* **123**, 163–178 (1998).
9. Thomason, J. J. *Can J. Zool.* **69**, 2326–2333 (1991).
10. Erickson, G. M. *et al.* *Nature* **382**, 706–708 (1996).
11. Wake, M. H. *Biol. Int.* **39**, 14–18 (2000).
12. Erickson, G. M. *Proc. Natl Acad. Sci. USA* **93**, 14623–14627 (1996).

particular functions of the sodium channel (Fig. 1; reviewed in ref. 6). The positively charged fourth transmembrane segment (S4) in each domain serves as the voltage sensor. The hydrophobic sixth transmembrane segments (S6) in the four domains form the wide inner lining of the pore. The preceding SS1/SS2 segments (also known as P loops) form the narrower, ion-selective outer lining of the pore. The short intracellular loop connecting domains III and IV inactivates the sodium channel within one or two milliseconds after it opens by folding over and blocking the inner mouth of the pore. This essential function is impaired in some inherited diseases of hyperexcitability, including periodic paralysis, cardiac arrhythmia and epilepsy (reviewed in ref. 7).

What are the three-dimensional structures of these functional elements of the sodium channel? Initial insights came from X-ray crystallography and nuclear magnetic resonance (NMR) analysis. The crystal structure of the pore region of a distantly related bacterial potassium channel⁸ defined a narrow outer pore formed by the P loops, and an inner pore surrounded by transmembrane segments that are related to the S6 segments of sodium channels. NMR analysis of the structure of the sodium channel's inactivation gate in solution revealed a rigid helical structure (an α -helix), preceded by two turns that position a key hydrophobic

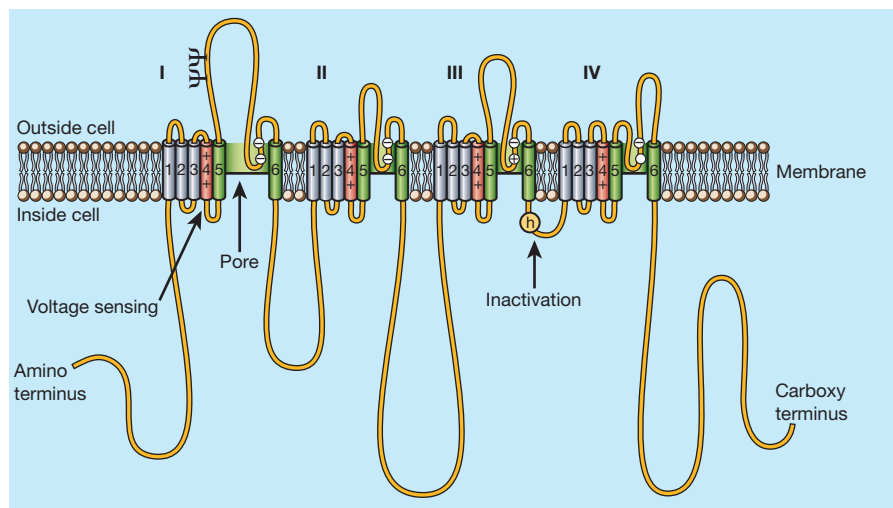


Figure 1 Two-dimensional molecular map of the α -subunit of the voltage-gated sodium channel, showing how it folds and crosses the plasma membrane. The α -subunit is composed of four homologous domains (I to IV), each of which contains six transmembrane segments (1 to 6). Cylinders represent α -helical transmembrane segments. Other protein segments are illustrated as straight lines, roughly in proportion to their length in the amino-acid sequence of the sodium channel. Red indicates the S4 transmembrane segments, which serve as voltage sensors. Green cylinders and shading indicate the S5 and S6 segments and the SS1/SS2 segment (P loop) between them, which together form the walls of the sodium-ion-conducting pore. Circles indicate the positions of amino acids (neutral, positively charged or negatively charged) that are important for ion conductance and selectivity. Ψ indicates sites of N-linked glycosylation (that is, where a carbohydrate group is attached). The circled 'h' (yellow) indicates the hydrophobic motif IFMT (isoleucine, phenylalanine, methionine, threonine) in the inactivation gate, which is thought to fold into and block the ion-conducting pore.

amino-acid sequence (isoleucine, phenylalanine, methionine, threonine) so that it can interact with and block the inner mouth of the pore⁹.

The structure of the complete sodium channel¹ now places these functional elements in a three-dimensional context (Fig. 2). Sato *et al.* determined this structure at 19 Å resolution by cryo-electron microscopy and image reconstruction. The images show four transmembrane masses arrayed symmetrically around a central axis that is perpendicular to the membrane. The central axis is densely stained with ions from the bathing solution, providing an image of the transmembrane pore and confirming the widely accepted idea that the four domains of the sodium channels are arranged roughly symmetrically around a central pore (Fig. 2, green). Remarkably, the central pore does not connect directly to the intracellular and extracellular bathing solutions. Instead, it splits into four branches that connect it to the bathing solutions at each end (Fig. 2, green).

As viewed parallel to the membrane surface, the sodium-channel protein is bell-shaped, with about 47% of its mass on the intracellular side of the membrane and 24% on the extracellular side (Fig. 2). These observations are in general agreement with expectations from two-dimensional models of the protein structure (Fig. 1). Much of the intracellular mass must be contributed

by the large amino-terminal and carboxy-terminal regions and the large intracellular loops connecting domains I, II and III. A prominent indentation in one side of the intracellular mass may hold the inactivation gate formed by the short intracellular loop connecting domains III and IV, as expected from the fact that this loop is accessible to enzymes, antibodies and chemical reagents (reviewed in ref. 6).

The most unexpected features of the structure are four peripherally located transmembrane pores, one in each domain (Fig. 2, red). What might these extra pores do? During gating, the voltage sensors (the S4 segments) of a sodium channel are predicted to move outwards across the membrane along a spiral path, thereby moving their positive gating charges across the membrane's electric field (reviewed in ref. 6). Outward movement and rotation of the S4 segments has indeed been detected with specific chemical labelling methods^{10–12}. The large outward movements observed for these transmembrane helices led to the proposal that these segments move through special, narrow-waisted gating pores in each domain, designed to allow the movement of substantial gating charge across the membrane with minimal physical motion¹⁰. The four narrow transmembrane pores now observed in each domain of the structure are prime candidates for these gating pores.

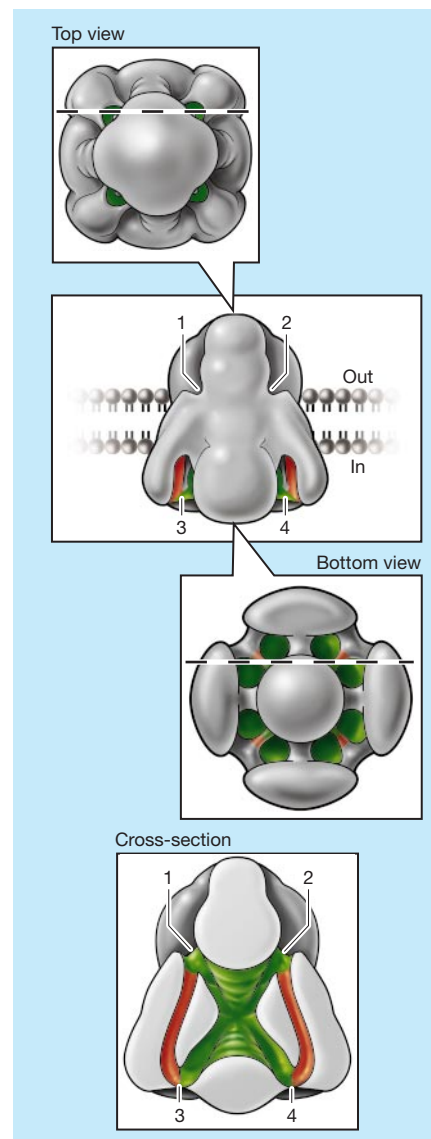


Figure 2 A three-dimensional representation of the voltage-gated sodium channel, based on Sato *et al.*'s images¹. Green indicates the central ion-conducting pore and its intracellular and extracellular connections to the bathing medium. Red indicates 'gating' pores, containing the S4 segments, and their connections to the central pore. Numbers 1 to 4 allow one to compare the orientations of the intact sodium channel (second image from the top) and the cross-section (bottom). The dashed lines show the position at which the cross-section was taken.

Surprisingly, these apparent gating pores are not linear, even though they are thought to accommodate the movement of a linear S4 segment across the membrane. Perhaps the structure observed in these images is a closed state in which transmembrane movement of the S4 segments is prevented. If so, voltage-dependent activation of the channel would require a conformational change in the gating pores to allow movement of the gating charge, as well as an outward translocation and rotation of the S4 segments

themselves. This would be a previously unrecognized step in the gating process.

This view of the sodium channel at 19 Å resolution whets our appetite for higher-resolution images. Analyses of such images, together with further structure-based studies of the channel's function, will help to answer many questions. For example, which molecular interactions underlie the gating process? How do the gating pores observed here participate in the overall regulation of sodium-channel function? And how does the structure of the central pore allow the selective passage of sodium ions? In the meantime, we can celebrate this latest achievement, which provides new insight into the structural basis of voltage-dependent gating, and holds great promise for the future. ■

William A. Catterall is in the Department of Pharmacology, School of Medicine, University of Washington, Seattle, Washington 98195-7280, USA.
e-mail: wcatt@u.washington.edu

1. Sato, C. *et al.* *Nature* **409**, 1047–1051 (2001).
2. Beneski, D. A. & Catterall, W. A. *Proc. Natl Acad. Sci. USA* **77**, 639–642 (1980).
3. Agnew, W. A., Moore, A. C., Levinson, S. R. & Raftery, M. A. *Biochem. Biophys. Res. Commun.* **92**, 860–866 (1980).
4. Noda, M. *et al.* *Nature* **312**, 121–127 (1984).
5. Goldin, A. L. *et al.* *Neuron* **28**, 365–368 (2000).
6. Catterall, W. A. *Neuron* **26**, 13–25 (2000).
7. Lehmann-Horn, F. & Jurkat-Rott, K. *Physiol. Rev.* **79**, 1317–1372 (1999).
8. Doyle, D. A. *et al.* *Science* **280**, 69–77 (1998).
9. Rohl, C. A., Boeckman, F. A., Scheuer, T., Catterall, W. A. & Klevit, R. *Biochemistry* **38**, 855–861 (1999).
10. Yang, N., George, A. L. & Horn, R. *Neuron* **16**, 113–122 (1996).
11. Cha, A., Snyder, G. E., Selvin, P. R. & Bezanilla, F. *Nature* **402**, 809–813 (1999).
12. Glauner, K. S., Manuzzi, L. M., Gandhi, C. S. & Isacoff, E. *Nature* **402**, 813–817 (1999).

Earth history

Sediments to planetary motion

Marie-France Loutre

Celestial mechanics has long been used as an aid for interpreting geological records. The compliment is being repaid through analysis of Mediterranean sediments.

Theories about the variation in Earth's orbit have proved helpful in setting the chronology for geological records of climate. But the reverse can also be true — such records can aid astronomers in computing Earth's orbital parameters at times in the remote past. That is the aim of Lourens *et al.*, as they describe on page 1029 of this issue¹. They have taken advantage of an exceptional sediment record, which can be related to insolation (the amount of solar radiation reaching Earth's upper atmosphere). They come to certain conclusions about the planet's rotational history over the past three million years, and create an astronomical timescale to be used in further research.

Climate records for all except the very recent past exist only as proxies, not as direct measurements. Such proxies are derived from tree rings or cores from various locations — the deep seas, lakes, continents, ice sheets or glaciers. The timescale is usually drawn from radioactive-decay techniques, allowing the development of a good chronology for up to a few hundred thousand years ago. Another procedure has also been devised to match (tune) the features in the proxy records to those of the Earth's orbital parameters or of the astronomically derived insolation (or both). Since the 1980s it has become clear that much climate variation stems in some way from changes in insolation driven by various astronomical factors — Earth's eccentricity (the shape of its orbit around the Sun), its obliquity (inclination of the axis of rotation), and its precession

(the planet's varying position on its orbit at the beginning of the seasons).

There are of course limitations to this tuning technique. The climate cycles in the data series must be recognized *a priori*, proceeding from causes to effects, and the climate record must contain at least some independent indication of its age (a so-called control point). The lag between the astronomically induced change and the climate response is usually assumed to be constant, although it may be different in (for instance) glacial and interglacial times. Finally, the accuracy of the chosen target curve — that is, the computed orbital parameters, insolation, or both — is limited, in particular by our imperfect knowledge of Solar System dynamics.

As climate records became longer and older, it became increasingly necessary to have highly accurate time series showing past variation in orbital parameters. Hence the proposal at a conference² some years ago that the geological record could be used to obtain estimates of past variation in orbital parameters, and that those estimates could be used to constrain the predictions based on celestial mechanics.

By the early 1990s, slightly different astronomical solutions became available for the period up to about several million years ago (see ref. 3 for review). One of them, known as the La90_(γ ,TD) solution⁴, accounts for two factors that influence Earth's orbit. The first, γ , is the dynamical ellipticity, which is affected by changes in the distribution of masses on and in the Earth (ice caps, for instance, or patterns of mantle convec-

tion). TD is tidal dissipation, which is related to the exchange of angular momentum between the Earth and the Moon, and results in changes in the Earth–Moon distance, in the spin velocity of the Earth, and in the velocity of the Moon's rotation about the Earth. Both dynamical ellipticity and tidal dissipation affect Earth's obliquity and precession, and have to be determined independently of astronomical theory. Each pair of parameters gives rise to one solution. Faced with a large number of slightly different solutions, it becomes necessary to try to identify the 'best' one.

In one earlier attempt, Lourens *et al.*⁵ correlated cycles seen in Mediterranean sediments over the past five million years with summer insolation derived from different astronomical solutions, one of which was La90_(1,1) — that is, with the present-day value for dynamical ellipticity ($\gamma = 0.9$ would mean 0.9 times today's value) and the present-day value for tidal friction (TD = 0 would mean no tidal dissipation). They looked at interference patterns between precession and obliquity that were recognizable in both the sediment data and computed insolation. A cross-spectral comparison between the data and summer insolation time series led them to conclude that the La90_(1,1) is the most accurate astronomical solution from a geological point of view — in other words, it results in the best fit with the geological record over the past five million years.

More recently, Pälike and Shackleton⁶ applied a method based on interference patterns between the precession and obliquity components derived from geological data, and the astronomical solutions, to identify small changes in tidal dissipation and dynamical ellipticity. They showed that these parameters are likely to have remained close to the present-day values for the past 25 million years.

In their new work¹, Lourens *et al.* present a comprehensive study to estimate dynamical ellipticity and tidal dissipation in the La90_(γ ,TD) solution between 2.5 and 3 million years ago. Their work is made possible by a high-resolution record of the ratio of titanium to aluminium from a sediment core in the eastern Mediterranean Sea. For complicated reasons to do with the balance between airborne and river input of material into the Mediterranean, this ratio provides a good estimate of humidity over northern Africa, which is controlled by summer insolation. The data series is tuned against an astronomical index computed for different La90_(γ ,TD) astronomical solutions — that is, the astronomical ages of the target are assigned to correlative peaks in the data series. The highest linear correlations (the optimal fits) between the target and the data series are obtained when the target is based on solutions ranging between La90_(1,0003,0) and La90_(0.9997,1). Such a high correlation is not unequivocal evidence

that the timescale is correct⁷. But it indicates a close agreement in the phase relationship between the series, and thus in the underlying frequencies.

From the optimal fit between their record and the astronomical index (La90_(1,0.5)), Lourens and colleagues conclude that the combined effects of dynamical ellipticity and tidal dissipation have, on average, been smaller over the past three million years than they are today. From this 'best fitting' astronomical solution, they can also determine the rate of change in the length of the day. Moreover, as dynamical ellipticity and tidal dissipation are parameters in the computation of obliquity and precession, the changes in these periods can also be estimated for the same time interval.

Although these new results¹ aren't necessarily in direct conflict with those of Pälike and Shackleton⁶, we obviously need a more precise picture of past variations in Earth's orbital parameters. First, Lourens and colleagues' results will have to be confirmed by similar studies of other climate records (which, at the same time, should reduce the uncertainty estimates on the results). Second, some assumptions, related for example to the lag between precession and obliquity and to the robustness of the control point, should be tested and justified. Third, the same analysis should be applied to a longer interval of time. Attempts⁸ to calibrate the proxy record much further back in time, for instance in the Early

Mesozoic around 200 million years ago, are giving promising results. Such attempts are mainly based on matching the proxy record to the variations in eccentricity, because eccentricity is much more stable through time than the other orbital parameters.

What of the future? Geological climate records are becoming available for longer periods of time. Different techniques are being used to help estimate the best fit between astronomical theory and geological records. And astronomical computations are continually being extended and improved. All in all, these complementary approaches mean bright prospects for improving understanding of both Earth's climate in the remote past and how the climate record can inform celestial mechanics. ■

Marie-France Loutre is in the Institut d'Astronomie et de Géophysique G. Lemaître, Université Catholique de Louvain, 2 Chemin du Cyclotron, B1348 Louvain-la-Neuve, Belgium.
e-mail: loutre@astr.ucl.ac.be

1. Lourens, L. J., Wehausen, R. & Brumsack, H. J. *Nature* **409**, 1029–1033 (2001).
2. Berger, A., Imbrie, J., Hays, J., Kukla, G. & Saltzman, B. (eds) *Milankovitch and Climate* 823–830 (Reidel, Dordrecht, 1984).
3. Berger, A. & Loutre, M. F. *Earth Planet. Sci. Lett.* **111**, 369–382 (1992).
4. Laskar, J., Joutel, F. & Boudin, F. *Astron. Astrophys.* **270**, 522–533 (1993).
5. Lourens, L. J. *et al. Paleoclimatology* **11**, 391–413 (1996).
6. Pälike, H. & Shackleton, N. *Earth Planet. Sci. Lett.* **182**, 1–14 (2000).
7. Shackleton, N. *et al. Paleoclimatology* **10**, 693–697 (1995).
8. Olsen, P. E. & Kent, D. V. *Phil. Trans. R. Soc. Lond. A* **357**, 1761–1786 (1999).

Ecology

Bagging the lag

Michael E. Hochberg and Arthur E. Weis

New techniques for analysing the dynamics of moth populations in captivity could have broad implications for biodiversity and conservation research in general.

Predicting changes in the size of natural populations has long been a goal of ecologists. But it is often far from simple to unravel the dynamics of populations. We rarely know all of the factors that affect them, let alone the many ways in which individual populations interact. Hence the importance of the paper on page 1001 of this issue¹, in which Bjørnstad and co-workers — using data collected from plastic boxes stocked with meal moths and their natural enemies — evaluate new ways of analysing population dynamics.

Theory predicts that populations are more likely to persist if their size is kept in check. In other words, there is a maximum number of individuals per population that can survive in a given environment. But it is difficult to detect such 'density dependence' and to estimate the extent to which it contributes to population stability^{2,3}. One prob-

lem facing ecologists is that many environmental factors impinge on population sizes, affecting density to different degrees. Such factors include abiotic variables, such as rainfall and temperature, and biotic influences, such as resources, competitors and natural enemies.

Is there any hope of disentangling the effects of this multitude of variables? If so, it would seem reasonable to start with long-term data collected from simple environments with only a few interacting species. Recent research⁴ indicates that data collected from natural populations can carry the characteristic imprints of an interaction (or 'coupling') between two populations. Bjørnstad *et al.*¹ go a step further, by showing how a different type of population signature — the lag between the introduction of a parasite and its effect on the density of a host population — may be associated with the coupling

of the dynamics of the host and parasite populations.

Bjørnstad *et al.*'s data come from their previous laboratory study⁵ of the moth *Plo-dia interpunctella*, a pest that feeds on stored products such as flour, nuts and dried fruit. Moths were cultured alone or with either the parasitoid wasp *Venturia canescens* (which lays its eggs in *P. interpunctella* larvae) or the *P. interpunctella* granulosis virus. In the absence of wasps or viruses, moth populations showed peaks in abundance with a period of one generation (so-called generation cycles). The viruses had little apparent effect when confined with the moths, but the wasps did, deepening the cycles' troughs. This broad difference formed the basis for Bjørnstad *et al.*'s present study¹.

When the dynamics of a population are dependent on its density, its size at one point in time may be influenced by its size and composition in the recent past. For instance, older *P. interpunctella* larvae eat younger ones, so low numbers of mature moths at one time may reflect the lagged effect of an earlier abundance of cannibals. Other interactions — including density-dependent parasitism — can cause further lags. Lags can be detected by determining whether the current population size can be predicted by its size at various times in the past. For example, when cultured alone, *P. interpunctella* larvae show three density-dependent population lags; the first of these reflects the state of the population a week previously, when the population density was dependent on competition for resources and the cannibalism of small larvae by large ones. Bjørnstad *et al.* refer to the number of lags as the 'dynamical dimension' (three, in this case).

Bjørnstad *et al.* dissect moth population dynamics with surgical precision. First, using a time-series method, they find that adding the virus to the moth population does not change the dynamical dimension, whereas adding the wasp parasite increases it by two lag terms. Second, the authors show that the number of adult moths decreases with past wasp abundance, but is not adversely affected by the past number of moth larvae that were infected with the virus. By untangling the dynamical dimensionality in each case, Bjørnstad *et al.* show that the moth and wasp populations, but not the moth and virus populations, are tightly coupled.

Next, to see more clearly the inner workings of the dynamical dimensions, the authors construct a simple population model, and use it to identify the processes that contribute to each lag. Interestingly, the parasitoid wasp not only increases the number of lags, but also qualitatively changes the lags. These changes, although complex, involve the addition of new terms reflecting trophic (feeding) interactions, and the mod-

ification of existing terms. Finally, Bjørnstad *et al.* apply a statistical method to their raw data, and confirm the findings based on the time-series method and the population model.

We do not yet know how useful Bjørnstad *et al.*'s methods will be in identifying the important variables in other systems, including natural ones. Theory predicts⁶ how the dynamics of a population will be affected by the strength of coupling between two species. But it is not yet clear whether coupling to more and more variables inexorably increases the number of lags.

However, on the empirical front, thousands of sets of population data exist⁷. Although it would take a herculean effort to analyse them all, breakthroughs may be in the offing. The same group previously showed⁸ that by increasing the depth of *P. interpunctella*'s artificial diet, the wasps' attack rate could be diminished, resulting in a weaker effect on the moth's population dynamics. Combining this system with the new analysis techniques will provide an opportunity to test whether varying a habitat parameter affects the strength of coupling of the system. The prediction here is that coupling between moth and wasp population dynamics should decrease as diet depth increases.

There are also broader implications. The group previously found⁵ that when both virus and wasp confront the moth together, the simple generation cycles found in the one- and two-species systems are thrown

out of whack: the three-species system exhibits transient cycles of longer periods, and eventually becomes extinct. If the imprints of one or both enemies on these cycles could be found, we would be a step closer to knowing whether Bjørnstad *et al.*'s techniques can be applied to biodiversity and conservation research, where one might want to know how a species that is harmless in one context can lead to the collapse of part of a community in another. These techniques could also be used to show how the use of two natural enemies as biological pest controls yields outcomes that are qualitatively different to the outcomes of using either enemy alone⁹.

Michael E. Hochberg is at the Institut des Sciences de l'Evolution, Université de Montpellier II, 34095 Montpellier, France.

e-mail: hochberg@isem.univ-montp2.fr

Arthur E. Weis is in the Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92697, USA.

e-mail: aeweis@uci.edu

1. Bjørnstad, O. N., Sait, S. M., Stenseth, N. C., Thompson, D. J. & Begon, M. *Nature* **409**, 1001–1006 (2001).
2. Woivod, I. P. & Hanski, I. *J. Anim. Ecol.* **61**, 619–629 (1992).
3. Strong, D. R. *Trends Ecol. Evol.* **1**, 39–42 (1986).
4. Turchin, P. *et al.* *Nature* **405**, 562–565 (2000).
5. Begon, M., Sait, S. M. & Thompson, D. J. *Nature* **381**, 311–315 (1996).
6. Tanner, J. T. *Ecology* **56**, 855–867 (1975).
7. NERC Centre for Population Biology, Imperial College. The Global Population Dynamics Database (1999). <http://cpbnts1.bio.ic.ac.uk/gpdd/>
8. Begon, M., Sait, S. M. & Thompson, D. J. *Proc. R. Soc. Lond. B* **260**, 131–137 (1995).
9. Murdoch, W. W. & Briggs, C. J. *Ecology* **77**, 2001–2003 (1996).

Turbulence

Go with the flow

Itamar Procaccia

Traditional devices for measuring turbulence have been unable to keep up with the latest developments in theory. But detectors derived from high-energy physics may narrow the gap between experiment and theory.

Turbulence is the chaotic and unpredictable motion of fluids flowing at high rates. It plays a major role in many processes from the environmental, for example cloud formation, to the technological, such as in industrial chemical reactors. Clearly, a deeper understanding of this phenomenon would be beneficial, and in recent years much progress has been made in the fundamental theory underlying turbulence. But the ability to measure turbulence experimentally has not advanced at the same rate, making it difficult to verify the theoretical developments.

On page 1017 of this issue, Eberhard Bodenschatz and collaborators¹ report an important technical improvement to the way in which turbulence is measured. This advance may make a decisive contribution to

bringing experimental turbulence research back on par with theory.

In their experiment, Bodenschatz and co-workers¹ modified a detector from Cornell's electron–positron collider. They used 'silicon strip' detectors to optically image tracer particles (tiny transparent beads) in turbulent water flow. Compared with previously available techniques, this method offers unprecedented time resolution of up to 70,000 frames per second. As a result, the researchers were able to measure the acceleration of the particles in turbulent water, discovering that it can reach 1,500 times the acceleration of gravity. The high time-resolution of the measurements indicates that the acceleration is highly intermittent, reflecting the complex structure of turbulent flow.

At present, the standard probe for turbu-



100 YEARS AGO

A simple workable, absolutely trustworthy system is still urgently wanted for the detection of criminals, and if the authoress of this book has succeeded she certainly deserves the thanks of all the Governments of Europe... It so happened that about seven years ago the reviewer came to the conclusion that the external ear ought to yield some clue to the relationship of man and ape, and of one race of man to another... To test the "criminal-mark" theory of Lombroso and many others, he examined the ears of more than 800 confirmed criminals, and of more than two thousand inmates of asylums for the insane, situated in parts of the country where he had already examined the ears of the sane. Altogether the ears of more than 40,000 people of different races and of different moralities, besides those of about 300 apes and anthropoids, were examined, but the total results of this elaborate investigation were almost entirely of a negative nature... If the reviewer's methods and observations are correct, the confirmed criminal's ear is the ear of the average inhabitant of Great Britain. Nor did the ears of the insane differ, on an average, from those of the people from which they were drawn, and if the authoress had carried her observations over a number of men of genius or of high ability, instead of drawing elaborate deductions from single observations, she would probably have arrived at a similar conclusion as to them.

From *Nature* 21 February 1901.

50 YEARS AGO

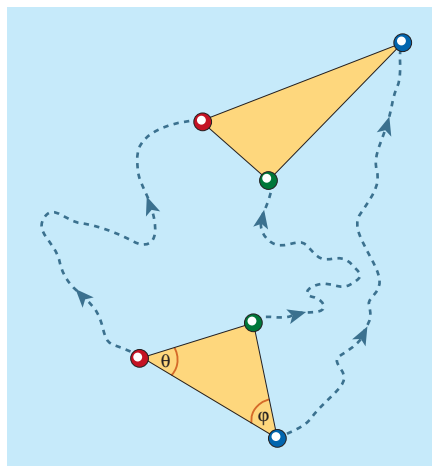
Miss Dorothea M. A. Bate, who died after a brief illness on January 13 at the age of seventy-two, was for more than fifty years one of the outstanding personalities at the British Museum (Natural History). When only seventeen, and with neither qualification nor encouragement, she started work in the Bird Room as a voluntary worker; but her interests lay chiefly in palaeontology in relation to the Recent fauna, rather than in the Recent fauna itself... During 1901–1902 Miss Bate explored the caves of Cyprus and made some notable discoveries, such as the remains of pigmy elephants, and soon extended her interest to cave deposits in Crete, the Balearic—where she discovered the unique 'antelope' *Myotragus*—Malta and Sardinia, working meticulously and earnestly and always alone.

From *Nature* 24 February 1951.

lence research is the hot-wire anemometer. This consists of a thin wire that is heated by the passage of an electrical current and is kept at a constant temperature by means of a feedback loop. The wire is placed at right angles to the average flow of a turbulent fluid. The fluid cools the wire and, because the cooling effect depends on the fluid's velocity — the faster the fluid flows, the more the wire is cooled — the velocity can be measured as a function of time. As would be expected for turbulent fluids, an erratic time series is obtained.

Hot-wire technology has progressed in recent years: for example, superconducting elements have been used to measure the velocities of cryogenic helium turbulence². But this technology gives only limited information on the structure of turbulence. Hot-wire anemometry is fundamentally limited: it can provide data only for a fixed point — in other words, for the point at which the wire is positioned. But theorists dream about measuring scale-dependent information from points that move with the flow — the so-called 'lagrangian' trajectories (Fig. 1).

Turbulent flow is irregular and disordered, with particles experiencing many different velocities. At large scales, the



chaotic flow of the fluid demonstrates the presence of high amounts of energy. But instability results in energy loss, which cascades nonlinearly down to smaller scales. This cascade continues down to the smallest scale where molecular friction comes in and energy is dissipated by viscosity. The most interesting regime to study covers the intermediate scales, where the presence of a constant energy flux from large to small scales establishes a statistical equilibrium inside

Figure 1 The lagrangian evolution of a group of three tracer particles. At any moment in time the trio defines a triangle that is fully determined by a scale R , the Euler angles of its orientation in space, and two internal angles. Theory focuses on 'statistically preserved structures' which are determined by the distribution on internal angles and the scale. These structures dominate the statistical theory.

the fluid. For theoretical calculations, the mean velocity of the fluid is subtracted from the turbulent flow by a 'galilean' transformation. Theorists focus on the lagrangian trajectories in the moving frame of the fluid, and are interested in measuring the physical phenomena in this lagrangian picture. The technique devised by Bodenschatz and co-workers brings us closer to this dream.

Hot-wire anemometry would be quite useless for turbulence research had it not been for the ingenuity of G. I. Taylor, who pointed out that when the mean velocity of the fluid is very high, the time series measured at a point can be considered as a spatial cut through the turbulent field. This 'Taylor frozen turbulence hypothesis' asserts that the turbulent field is swept through the probe faster than the field can appreciably change, so the wire is taking a one-dimensional 'snapshot' of the velocity field. This has been the basis for analysis of turbulence data for decades, but besides being only approximately true, recent theoretical work has drawn attention to essential features of turbulence that are totally missed by measuring one-dimensional cuts (Box 1).

But Bodenschatz and co-workers offer new possibilities in following the detailed motion of fluid particles. This progress promises an exciting and fruitful interaction between theory and experiments in turbulence in the coming years. ■

Itamar Procaccia is in the Department of Chemical Physics, The Weizmann Institute of Science, Rehovot 76100, Israel.

e-mail: itamar.procaccia@weizmann.ac.il

1. La Porta, A., Voth, G. A., Crawford, A. M., Alexander, J. & Bodenschatz, E. *Nature* **409**, 1017–1019 (2001).
2. Tabeling, P. *Phys. Rev. E* **53**, 1613–1621 (1996).
3. Bernard, D., Gawedcki, K. & Kupiainen, A. *J. Stat. Phys.* **90**, 519–569 (1998).
4. Gat, O., Procaccia, I. & Zeitak, R. *Phys. Rev. Lett.* **80**, 5536–5539 (1998).
5. Frisch, U., Mazzino, A. & Vergassola, M. *Phys. Rev. Lett.* **80**, 5532–5535 (1998).
6. Arad, I. & Procaccia, I. in *Proc. IUTAM Symp. Geom. Stat. Turb.* (ed. Kambe, T.) 175–184 (Kluwer, Dordrecht, 2001).
7. Celani, A. & Vergassola, M. *Phys. Rev. Lett.* **86**, 424–427 (2001).
8. Arad, I., Biferale, L., Celani, A., Procaccia, I. & Vergassola, M. (submitted).

Box 1 A problem to be solved

One of the theoretical problems that can be verified experimentally using the techniques developed by Bodenschatz and co-workers¹ involves an important aspect of turbulence: its role in dispersing contaminants such as smoke or radioactive elements in the air and pollutants in the ocean. The concentration of such a contaminant at a space point r at time t is denoted as $T(r, t)$. One may be interested in the expected value of the field at some point $T(r_3, t)$ given the measurement of the concentration at two other points r_1 and r_2 at the same time. Such expected values are related to 'correlation functions'. In this example, we have a third-order correlation function which is denoted as $\langle T(r_1, t) T(r_2, t) T(r_3, t) \rangle$ where the averaging is over time.

Central to the statistical theory of turbulence is the finding that such correlation functions show a remarkable property called 'scaling'. This means that if we increase all the distances between the measurement points r_1 , r_2 and r_3 by a given factor λ , then the value of the correlation function changes by a multiplicative factor λ^{ζ_3} , where ζ_3 is a characteristic exponent. If we consider higher-order correlation functions between four, five or n points, they have the same property but with exponents ζ_n that depend on the order of the correlation function. These scaling exponents ζ_n are believed to be universal characteristics of the small-scale structure of the turbulent velocity field, reflected in the structure functions of advected contaminant. Their theoretical calculation is much sought after in fundamental turbulence research.

In recent years there has been a fundamental shift in the theoretical approach to such characteristics of turbulence³. It turns out that the nature of these exponents is related to subtle geometrical properties of groups of lagrangian trajectories of tracer particles^{4,5}. For example, to understand the exponent ζ_3 one needs to focus on the dynamics of three tracer particles. Obviously, at any point in time three tracer particles define a triangle, which in turn is fully characterized by one length scale R (say the geometric mean of the lengths of its sides) and two angles, say θ and ϕ . When the three tracer particles are advected by the turbulent velocity field (see Fig. 1 on page 1017 for the lagrangian trajectory of one such particle), the scale R of the defined triangle and its shape (angles) change continuously (Fig. 1). If we rescale the triangle to size $R=1$, the dynamics become a trajectory in the space of shapes⁶, the space of all triangles of size 1 (conveniently characterized by two angles θ and ϕ). These dynamics have equilibrium distributions given by $\rho(\theta, \phi)$. The deep and surprising new statement that can be made is that the three-point statistics are dominated by trajectories in which the change in R is compensated by a change in shape so that $R^{\zeta_3} \rho(\theta, \phi)$ remains invariant⁷.

Such 'statistically preserved structures' are crucial for the statistical theory, as they obviously come to dominate the statistics⁸. Indeed, exponents such as ζ_3 can be understood as the rescaling exponents characterizing precisely such special distributions: the correlation function $\langle T(r_1, t) T(r_2, t) T(r_3, t) \rangle$ is proportional to $R^{\zeta_3} \rho(\theta, \phi)$, where R is the geometric mean of the distances between the points $r_1 \dots r_3$. Of course, the same ideas apply to any order correlation function or structure functions with the appropriate shape dynamics, and they quickly become the new language used to discuss scaling phenomena in turbulent transport^{7,8}.

L.P.

Obituary

Tom Kilburn (1921–2001)

By 1947 the theory of general-purpose, stored-program computers was familiar to several research groups on both sides of the Atlantic. However, the lack of suitable high-speed memory devices stopped them turning theory into practice. To quote Nathaniel Rochester of IBM, writing in the premier US electronics journal (*Proc. IRE*, p. 374; April 1950): "The most difficult problem in the construction of large-scale digital computers continues to be the question of how to build a memory, and the few papers written do not reflect the greatness of the effort which is being exerted." It was in this context that Tom Kilburn began working for his PhD at the University of Manchester. His task was to perfect a new form of digital memory. The starting point was an observation by F. C. Williams of the stored patterns of electrostatic charge inside a cathode-ray tube.

By the autumn of 1947 Kilburn had constructed a practical random-access memory device, later known as the Williams–Kilburn tube. Williams and Kilburn decided to incorporate the tube into a small-scale experimental computer, "to subject the memory system to the most searching tests possible". This small-scale computer, sometimes called the 'Baby', first ran a program on the morning of Monday 21 June 1948. The event was described in a letter to *Nature* published later that year (162, 487; 1948).

The 1948 Manchester computer contained only a few words (128 bytes) of storage. It was built from war-surplus electronic components, the rudimentary input being via push-buttons recycled from the radio-channel selector panels of Spitfire aircraft. Yet this small machine provided the first convincing demonstration of the stored-program principle, which is the basis of every modern computer. From this prototype, the British company Ferranti Ltd developed the world's first commercially available computer, known as the Ferranti Mark I. The first production model was delivered in February 1951, marginally ahead of the much larger UNIVAC 1 computer developed by Eckert and Mauchly in the United States.

Designing and building computers is a team activity combining hardware and software skills. When Williams moved on to other things, Kilburn built up a research team at the University of Manchester that completed a total of five increasingly powerful computers between



Designer of the first stored-program computer

1948 and 1975. The fourth of these, Atlas, was the world's most powerful computer when it came into operation in December 1962. Although this pre-eminence lasted but a few months, several innovative concepts that remain in use today were pioneered in Atlas, including virtual memory. Kilburn (seen here at the keyboard of Atlas) was particularly proud that all five designs led to industrial derivatives. Fruitful links with companies such as Ferranti and ICL meant that only one of the projects required an initial injection of public funding from Britain's Science Research Council. In 1964 the Manchester group evolved into the department of computer science, with Kilburn at its head.

But what of the man himself? Born in the no-nonsense county of Yorkshire, Kilburn was christened plain 'Tom' (not Thomas). Encouraged at school by an outstanding mathematics teacher, Kilburn won a scholarship to the University of Cambridge in 1940. He graduated with first-class honours in a two-year wartime mathematics course, and was given a quick introduction to electronics in London and sent to work at the Telecommunications Research Establishment, Malvern. Here he joined Williams, who headed a special engineering group solving electronic circuit problems for radar applications. Although originally a mathematician, Kilburn soon came to regard himself as an engineer. As with many young British

scientists at this time, his wartime experience in radar made him determined to see projects through to completion. Kilburn's enthusiasm for implementation became the hallmark of computer science at Manchester. Kilburn was accorded the scientific accolade of fellowship of the Royal Society in 1965.

Kilburn was a modest individual who actively shunned publicity. In 1968 he was asked why computer science textbooks seldom mentioned the early pioneering work at Manchester. Kilburn paused, and then replied mysteriously: "Because those who need to know, do know." Such isolationism was sometimes the despair of his more gregarious colleagues, but Kilburn's ability to remain focused on a problem enabled him to lead his team from the front.

In a field of scientific endeavour characterized by transitory, ad hoc developments, it is not easy to explain in simple terms the contributions to computer design made at Manchester between 1948 and 1975. If there was an underlying theme, it was the pursuit of hardware assistance for efficient scientific computation. Although some of Kilburn's engineering innovations were soon made obsolete by subsequent technologies (the advent of semiconductors, for instance), other ideas have stood the test of time. Among them are hardware assistance for accessing structured data such as arrays (the index register, 1949); hardware implementation of floating-point arithmetic (1954); hardware management of heterogeneous storage systems (associative page-address translation, 1962); and hardware management of local scalars in block-structured high-level languages (the name store, 1974).

In June 1948, the information revolution took its first tentative steps. By the end of 1949 there were probably still only four prototype computers coming into hesitant operation anywhere in the world — two in Britain (at the universities of Cambridge and Manchester), one in the United States (at the Eckert–Mauchly Computer Corporation, Philadelphia) and one in Australia (at the Commonwealth Scientific and Industrial Research Organisation, Sydney). Like its designer, the first Manchester computer was a creation of few words in a new world. Kilburn died on 17 January, aged 79, but his 'few words' will continue to resonate widely.

Simon Lavington

*Simon Lavington is in the Department of Computer Science, University of Essex, Colchester CO4 3SQ, UK.
e-mail: lavis@essex.ac.uk*

Muscle strength in overwintering bears

Unlike humans, bears retain their muscle tone when moribund for long periods.

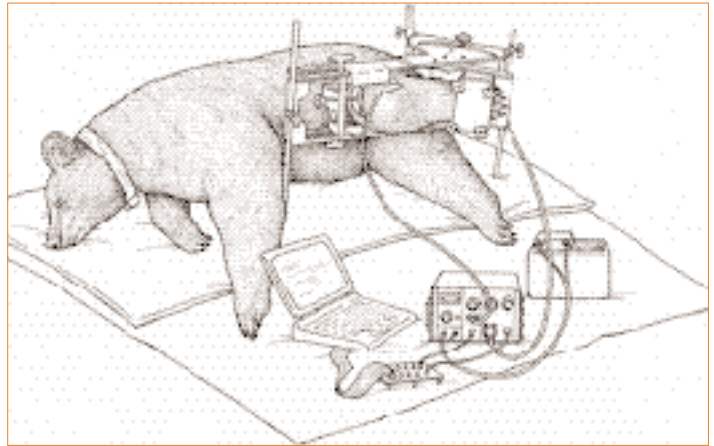
Black bears (*Ursus americanus*) stay inside their winter den for five to seven months of the year, during which time their body temperature drops to about 4 °C below normal¹ and they do not eat, drink, urinate, defecate or show any other perceptible activity². Although inactivity in humans, for example as a result of confined bed rest, weightlessness or limb immobilization, leads to atrophy of skeletal muscle, loss of muscle tone and impaired strength, we show here that the black bear does not suffer a similar deterioration. We find that overwintering black bears lose less than 23% of their strength over 130 days — unlike humans, who are weakened by a predicted 90% strength loss over the same period.

In rodents and humans experiencing muscle atrophy resulting from inactivity or starvation, there is a reduction in the number and size of muscle cells (fibres)³, in the ratio of slow-oxidative to fast-glycolytic fibres⁴, and in the myofibrillar protein content⁵. We have investigated how these parameters are affected in overwintering black bears by taking muscle biopsies from dened bears during early and late winter. Remarkably, we found no loss of skeletal-muscle cell number or size. In addition, some muscles retained their protein content and oxidative capacity completely, whereas others showed a slight reduction in the ratio of the two metabolic types of fibre and in their protein content⁶.

In view of these findings that bear muscle shows none of the typical signs of disuse atrophy and indicating that bears are likely to retain their locomotor prowess through the winter, we tested whether their muscle strength was affected by prolonged starvation and immobility. We monitored hind-leg flexion of bears in the field during late autumn, just after denning, and again in early spring before they emerged from their dens (Fig. 1). Muscle strength was measured using a modification of a non-invasive force-assessment technique that had been previously validated in a hospital setting on humans for diagnosing neuromuscular disorders, for assessing weakness due to disease progression, and for evaluating therapy⁷. We were able to quantify the strength of the dorsiflexor muscle (tibialis anterior) of wild bears in their natural setting by measuring muscle contraction in response to supramaximal stimulation of the common perineal nerve⁸.

Human strength tested on a similar apparatus registers typically as around 0.54 newton metres (Nm), which is comparable to the strength predicted for bears by

Figure 1 Non-invasive system for measuring strength of muscles by force assessment in a sleeping bear. Components consist of a stabilizing metal brace that holds the bear's leg in a secure horizontal position; a foot plate with force transducer to detect the evoked torque produced by the electrically stimulated tibialis anterior; a hardware device for nerve stimulation, signal amplification and signal conditioning, plus a computer for stimulus delivery, equipped with data-acquisition software for recording, analysing and displaying all signals simultaneously, together with a battery power-source.



extrapolation to their hibernation muscle temperature⁹. Muscle force measurements in lower-limb extensor and flexor muscles of humans under conditions of confined hospital-bed rest or simulated space flight show that there is a daily loss of strength of about 0.7% as a result of limb unloading¹⁰. If strength loss as a result of atrophy were to occur at this rate in bears, a decrease in peak torque force to about 0.0675 Nm (90% loss of strength) would be expected after 130 days of inactivity (Fig. 2). But surprisingly, even though we found a significant decrease in torque force, bear strength remained at around 0.35 Nm under these conditions, corresponding to a loss in strength of only 23% (Fig. 2).

Skeletal muscle protein and strength can be conserved by using alternative sources of

protein, by recycling urea nitrogen back into protein synthesis, or by rhythmically stimulating the muscles. Hibernating bears may therefore be able to retain their strength by synthesizing new amino acids and protein from urea nitrogen¹¹, or by shivering and undergoing isometric muscle contraction through the winter, or by drawing on labile protein reserves such as visceral smooth muscle and extracellular matrix. Understanding these processes in hibernating bears may provide new insight into treating muscle disorders and into the effects of prolonged hospital-bed confinement, antigravity and long-distance space travel on humans.

Henry J. Harlow*, Tom Lohuis*, Thomas D. I. Beck†, Paul A. Iaizzo‡

*Department of Zoology and Physiology, University of Wyoming, Laramie, Wyoming 82071, USA
e-mail: hharlow@uwyo.edu

†Colorado Division of Wildlife, 23929 Colorado Road U, Dolores, Colorado 81323, USA

‡Departments of Anesthesiology and Physiology, University of Minnesota, Minneapolis, Minnesota 55455, USA

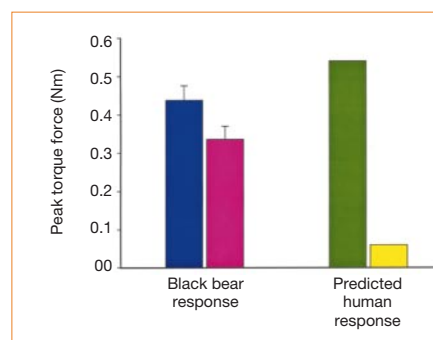


Figure 2 Muscle strength in hibernating black bears. Left, peak torque force (in newton metres, Nm) of the tibialis anterior in six bears tested in response to a multiple-pulse stimulus to the perineal nerve. Bears were tested at their den site in the autumn (blue bar) and again in the spring (pink bar); vertical lines are 1 s.e.m. Right, peak torque force (Nm) of ankle dorsiflexion in healthy humans determined using a similar apparatus and protocol. Green bar, predicted strength at a leg temperature similar to that of the bears; yellow bar, predicted strength after 130 days of atrophy at 0.7% strength loss per day.

1. Nelson, R. A. et al. *Int. Conf. Bear Res. Manag.* **5**, 284–290 (1983).
2. Nelson, R. A. et al. *Am. J. Physiol.* **224**, 491–496 (1973).
3. Nicks, D. K., Beneke, W. M., Key, R. M. & Timson, B. F. *J. Anat.* **163**, 1–5 (1989).
4. Musacchia, X. J., Steffen, J. M. & Deavers, D. R. *Aviat. Space Environ. Med.* **54**, 1015–1020 (1983).
5. Booth, F. W. & Kelso, J. R. *Pflugers Arch.* **342**, 231–238 (1973).
6. Tinker, D. B., Harlow, H. J. & Beck, T. D. I. *Physiol. Zool.* **71**, 414–424 (1998).
7. Sigg, M. D., Hasinoff, I. K. & Iaizzo, P. A. *Progr. Anesthesiol.* **XII**, 127–152 (1998).
8. Brass, T. J., Loushin, M. K. H., Day, J. W. & Iaizzo, P. A. *J. Med. Eng. Tech.* **20**, 67–74 (1996).
9. Quinlan, J. G., Iaizzo, P. A., Lambert, E. H. & Gronert, G. A. *Musc. Nerve* **12**, 119–125 (1989).
10. Berg, H. E., Larsson, L. & Tesch, P. A. *J. Appl. Physiol.* **82**, 182–188 (1997).
11. Nelson, R. A., Jones, J. D., Wahnner, H. W., McGill, D. B. & Code, C. F. *Mayo Clin. Proc.* **50**, 141–146 (1975).

Public health

Particulate emission from biomass combustion

Mixing dried human sewage sludge with pulverized coal is being evaluated as a way to reduce carbon dioxide emissions from traditionally coal-fired power plants, as well as to overcome sewage-disposal problems. Here we investigate the effects of inhaling particles emitted from the combustion of this mixture and find that these cause significantly more lung damage in mice than do particles from coal alone, probably because of their zinc content. Our findings indicate that the use of dried municipal sewage sludge as a 'green' (CO₂-neutral) replacement fuel should be considered with caution.

Airborne particulate matter is associated with acute respiratory distress in humans¹. Suggested causes² of the lung damage caused by these particles include their composition — for example, they may contain soluble transition metals such as copper, iron, vanadium, nickel or zinc — their acidity, and their ultrafine size (some particles are less than 0.1 µm in diameter). These properties are all features of airborne particulate matter resulting from the co-combustion of pulverized coal and biomass, including dried municipal sewage sludge (MSS).

Combustion of pulverized coal (from the University of Stuttgart) and of MSS/coal mixtures was performed in a semi-industrial-scale, 500-kW downfired pulverized-fuel combustor³. MSS had been processed (Swiss 'Combi' process) to produce 2–4-mm pellets, which were then pulverized and mixed (20% thermal, 50% mass load) with coal. Sampled particulate matter was resuspended and diluted⁴ to a sufficiently low concentration to allow the effects caused by different particle properties to be assessed on exposed mice. Mice were exposed for only 1 hour each day for 24 consecutive days to a low dose of 1,000 µg m⁻³ and to a high dose of 3,000 µg m⁻³ (the atmospheric dose of particulate matter less than 10 µm in diameter currently allowed by the United States Environmental Protection Agency is 150 µg m⁻³ averaged over 24 hours).

Figure 1 indicates that co-combustion of MSS/coal ash produces airborne particulate matter that has a significant effect on lung permeability in mice, increasing with dosage, which is greater than the effect of either air on its own or a high dose of particles produced by burning only coal ash, which appears to be relatively benign. Cell counts in broncho-alveolar lavage fluid in our mice decreased after exposure to particulate matter from MSS/coal, compared to ambient air and coal ash alone, which again appears relatively benign (Fig. 1). (Details of measurement techniques and the use of

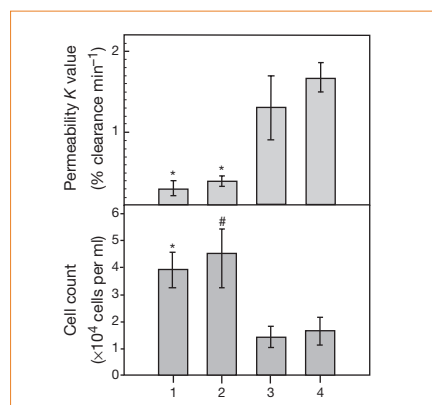


Figure 1 Effect of inhaled particulates on lung permeability (top) and broncho-alveolar lavage fluid cell counts (bottom). Thirty-two specific, pathogen-free mice (strain C57BL/6) were randomly divided into four groups (weight, 25.0 ± 4.3 g; n = 8), housed at 8 per cage under a 12-h light/12-h dark cycle at the Arizona Health Science Center animal facility (approved by the American Association for the Accreditation of Laboratory Animal Care, to ensure proper treatment of animals), and fed a standard chow diet and tap water. Results were analysed 24–30 h after the final exposure to airborne particulate matter (see text). Cells were counted from the first 3 of 6 lavages (1 ml sterile isotonic saline) of the lungs. In the top bar graph, the asterisk symbol indicates statistical significance compared with exposure to particles from the coal/sewage ash of $P < 0.0001$; in the bottom graph, the asterisk and hash indicate statistical significance compared with exposure to coal/sewage ash particles of $P < 0.0005$ and $P < 0.0001$, respectively. Error bars indicate ± s.d. Bars: 1, controls; 2, values corresponding to coal-ash exposure dose of ~3,000 µg m⁻³; 3 and 4, values for coal/sewage-ash exposures of ~1,000 µg m⁻³ and ~3,000 µg m⁻³, respectively.

anaesthetic and paralytic drugs to minimize discomfort and pain in the mice can be obtained from M.L.W.) We found no significant differences in other pulmonary functions (pulmonary compliance or resistance). Increased lung permeability enhances access by a toxin to the lung interstitium and pulmonary circulation, correlating with the extent of pathological lung injury.

We measured the distribution in particle sizes to which the mice were exposed by using a Berner low-pressure impactor, and recorded differences in mass and concentration of certain elements. Mass and size distributions for particles produced by coal/MSS ash mixtures and by coal ash on its own are quite similar, although the coal ash generates a slightly greater fraction of nanometre-sized particles. We conclude that the increased toxicity of the MSS/coal is not due to a difference in the size of the particles it produces on combustion.

Instead, the toxic effects seem to be associated with the presence of zinc in the particles. Combustion of MSS together with coal increases the concentration of zinc in the resulting airborne particulate matter in the respirable-size particles (diameters between 0.3 and 2.5 µm) from 2 to 14 wt% and in ultrafine particles (diameter smaller than 0.1 µm) from 4 to 11 wt%. MSS also produces

10 wt% cobalt in the ultrafines. Selenium, arsenic, lead and vanadium were present in similar amounts in particles from MSS/coal and from coal, whereas more iron was found in particles from coal burned alone. We therefore believe that zinc is the culprit metal that causes an increase in a measurable precursor for lung damage, a conclusion that is supported by other measurements⁵ of lung inflammation after intratracheal injection.

The replacement of coal by MSS in power stations is being considered as one option to help meet the requirements of the Kyoto Protocol and to replace landfill disposal of MSS. Our results, indicating that the toxicity to mice of inhaling particles from coal/MSS co-combustion is greater than from coal alone, suggest that the environmental advantages may be tempered by the risk to respiratory health.

Art Fernandez*, Sheldon B. Davis*, Jost O. L. Wendt*, Roberta Cenni†, R. Scott Young‡, Mark L. Witten‡

*Department of Chemical and Environmental Engineering, University of Arizona, Tucson, Arizona 85721, USA

e-mail: wendt@u.arizona.edu

†Institut für Verfahrenstechnik und Dampfkesselwesen, Universität Stuttgart, D-70569 Stuttgart, Germany

‡Lung Injury Laboratory, Department of Pediatrics, Arizona Health Sciences Center, Tucson, Arizona 85721, USA

1. Dockery, D. W. et al. *N. Engl. J. Med.* **329**, 1753–1759 (1993).
2. Lighty, J. S., Veranth, J. M. & Sarofim, A. F. *J. Air Waste Manag. Assoc.* **50**, 1565–1618 (2000).
3. Cenni, R., Frandsen, F., Gerhardt, T., Spliethoff, H. & Hein, K. R. G. *Waste Manag.* **18**, 433–444 (1998).
4. Fernandez, A., Wendt, J. O. L., Cenni, R., Young, R. S. & Witten, M. L. *J. Air Waste Manag. Assoc.* (submitted).
5. Adamson, I. Y. R., Prieditis, H., Hedgecock, C. & Vincent, R. *Toxicol. Appl. Pharmacol.* **166**, 111–119 (2000).

Palaeontology

An Early Cretaceous pellet

We have discovered a mass of fossil bones from four juvenile birds at Las Hoyas in Cuenca, Spain, which show signs of having been digested. To our knowledge, this rare finding of an Early Cretaceous fossil of an apparently regurgitated pellet provides the first evidence that Mesozoic birds were prey animals.

The new fossil assemblage contains four juvenile birds in an area smaller than 23 cm² (Fig. 1a,b). The two largest individuals (1 and 2; shown in dark and light grey, respectively, in Fig. 1) differ from one another in their tail morphology. In bird 2, the first seven caudal vertebrae are followed by a large pygostyle; in bird 1, the tail is composed of at least 12 free caudal vertebrae (Fig. 1c). The femur of bird 1 is 5–10% longer than that of bird 2, whereas the tar-

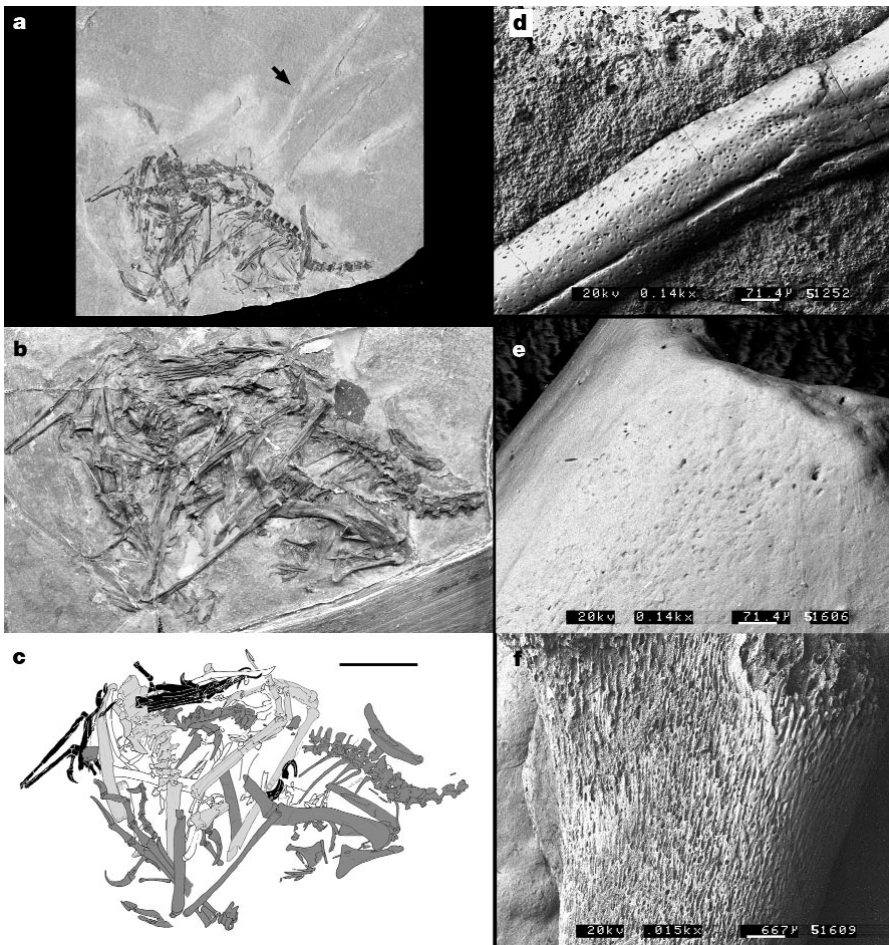


Figure 1 Fossil pellet from the Lower Cretaceous of Las Hoyas (Cuenca, Spain). **a**, Bird assemblage from Las Hoyas (specimen LH 11386, Museo de las Ciencias de Castilla-La Mancha, Cuenca, Spain), found by a local collector, D. A. Díaz Romeral) showing feather impressions (arrow). **b**, Acid preparation of the specimen LH 11386. **c**, Camera lucida drawing of the bird assemblage from Las Hoyas. The most complete individual (bird 1; dark grey) shows preservation of almost the entire vertebral column in articulation and multiple portions of the appendicular skeleton in varying degrees of articulation. The second individual (bird 2; light grey) is identifiable from sections of its pelvis, sacral and caudal series, and hindlimbs. The remaining two individuals (birds 3 and 4; black) are significantly smaller than the first two. These specimens can be identified by only their hindlimb elements; at least three small metatarsi are present in the assemblage. Individuals 1 and 2 are of roughly the same size but significantly larger than individuals 3 and 4. For example, the tarsometatarsi of individuals 3 and 4 are about 60% and 50% shorter than those of individuals 1 and 2, respectively. **d**, Scanning electron micrograph (SEM) of specimen LH 11386 revealing surface damage (pitting and edge rounding) by digestion. Scale bar, 71.4 μm . **e**, SEM of a modern avian bone digested by a jackal (courtesy of P. Andrews). Crocodiles produce very extreme digestion grades^{8,10}, higher than jackals and other canids⁵. Scale bar, 71.4 μm . **f**, SEM of an undigested modern juvenile bird bone (P. Andrews), showing absence of rounding and the characteristic porous and fibrous surface of juveniles. The specimen shown in **f** is larger than that in **d** or **e** and the magnification is reduced (scale bar, 667 μm) so that structural features can be compared.

sometatarsus of bird 1 is less than 85% the length of that of bird 2. The smaller individuals (birds 3 and 4) are comparable in size and have similar anatomical traits. The four individuals show a comparable degree of ossification. Morphological and size data suggest that at least three different species are present in the fossil.

This aggregation is peculiar to the Las Hoyas site^{1,2}, which has yielded more than 200 terrestrial vertebrate specimens that were usually isolated, fully articulated and complete. Results from oxygen and carbon isotope analysis of the Las Hoyas lacustrine limestones are in the range typical for sediments deposited in a still body of water³. Under these conditions, it would be unlikely that the four immature birds of this bone

assemblage perished far apart and were then buried together in such a small area.

There are two more parsimonious explanations. The individuals could have been living in a nest that was washed into the lake, or their association could have resulted from predation as individuals and regurgitation by the predator combined into a pellet. The first of these two non-random explanations assumes nest parasitism, because the four individuals probably belong to three different bird species. Such a behaviour is unknown for any basal bird and cannot be confidently inferred from our knowledge of their closest relatives.

The idea that the fossil is a result of predation is supported by the corroded aspect of some articular ends, surface pitting and

edge rounding of many of the bones (Fig. 1d), which look like digested bones from modern birds⁴ (Fig. 1e) and mammals⁵. These features differ from those expected from any other type of biological or geological erosion^{6,7}. The surface of bones from juvenile birds is also porous during periosteal formation, but this form of pitting is more fibrous and can be distinguished from that seen in digested bones (Fig. 1f).

Comparison with digested prey from extant predators indicates that the digestion of the Las Hoyas bone assemblage was extensive, although not as thorough as prey eaten by crocodiles⁸. The nearly intact anatomical connections of the bones contrast with the small size of bone fragments in dinosaur coprolites⁹. Impressions of complete feathers surrounding the assemblage of bones (Fig. 1a) resemble the feathers often seen enveloping modern pellets. These observations, together with the apparent absence of faecal ground mass, indicate that this multispecific aggregation of immature bird bones is a regurgitated pellet.

Identifying the predator responsible for this pellet is a problem. On the basis of the existing (and expected) fossil record from Las Hoyas, likely candidates could be large fish, early mammals, lizards, crocodiles (goniopholids), pterosaurs and dinosaurs (including birds), although some can be confidently excluded. Fishes and mammals regurgitate loose masses of bones, and the lizards and birds found at Las Hoyas are of comparable size to the pellet structure reported here. Goniopholids, close relatives of modern crocodiles, would probably have regurgitated more completely digested bones. The most likely predators to have produced the pellet are small, non-avian, theropod dinosaurs or pterosaurs that hunted different prey, swallowed them whole and then regurgitated the indigestible remains, much as owls do today.

José L. Sanz*, **Luis M. Chiappe†**,
Yolanda Fernández-Jalvo‡,
Francisco Ortega*, **Begoña Sánchez-Chillón‡**, **Francisco J. Poyato-Ariza***,
Bernardino P. Pérez-Moreno*

*Unidad de Paleontología, Departamento de Biología, Universidad Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain
e-mail: jlsanz@inves.es

†Natural History Museum of Los Angeles County, 900 Exposition Boulevard, Los Angeles, California 90007, USA

‡Museo Nacional de Ciencias Naturales, José Gutiérrez Abascal 2, 28006 Madrid, Spain

1. Pérez-Moreno, B. P. *et al.* *Nature* **370**, 363–367 (1994).
2. Sanz, J. L., Fregenal-Martínez, M. A., Meléndez, N. & Ortega, F. in *Palaeobiology II* (eds Briggs, D. E. G. & Crowther, P. R.) (Blackwell Scientific, Oxford, in the press).
3. Poyato-Ariza, F. J., Talbot, M. R., Fregenal-Martínez, M. A., Meléndez, N. & Wenz, S. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **144**, 65–84 (1998).
4. Bochenki, Z. M. *et al.* *J. Archaeol. Sci.* **25**, 425–433 (1998).
5. Andrews, P. *Owls, Caves and Fossils* (Natural History Museum Publications, London, 1990).

6. Davis, P. *Int. J. Osteoarchaeol.* **7**, 388–401 (1997).
7. Behrensmeier, A. K. in *Taphonomy: Releasing the Data Locked in the Fossil Record* (eds Allison, P. A. & Briggs, D. E. G.) 291–335 (Plenum, New York, 1991).
8. Fisher, D. C. *Paleobiology* **7**, 262–275 (1981).
9. Chin, K., Tokaryk, T. T., Erickson, G. M. & Calk, L. C. *Nature* **393**, 680–682 (1998).
10. Andrews, P. & Fernández-Jalvo, Y. *Nature* **393**, 629–630 (1998).

Molecular biology

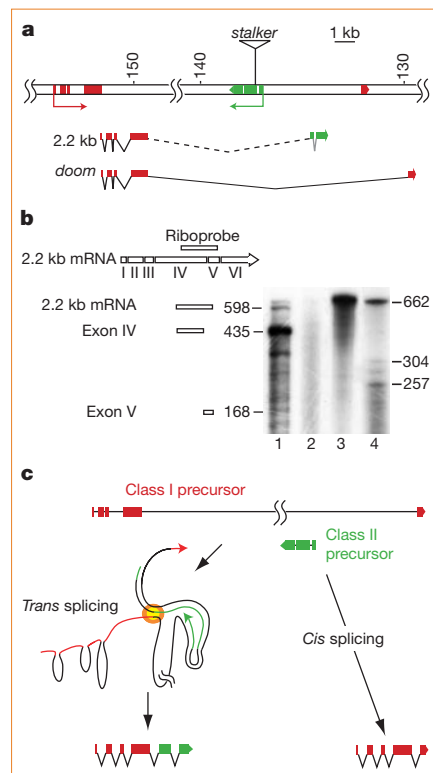
Protein encoding by both DNA strands

All the evidence so far points to a gene's protein-coding information being contained in only one of its two DNA strands, with this strand serving as a template for transcription of the precursor RNA that is eventually translated into protein¹. Here we present structural evidence showing that the protein-coding information of the *modifier of mdg4* (*mod(mdg4)*) gene of the fruitfly *Drosophila* is provided by both of its complementary DNA strands, and not by just one. This novel organization means that RNA precursors generated from two DNA templates need to be joined subsequently into a single messenger RNA, a surprising feature that raises new questions regarding genome complexity and evolution.

The *mod(mdg4)* proteins of *Drosophila* are required during the fly's development and probably help to establish or maintain chromatin structure^{2–4}. The genomic sequence of this locus has been determined⁵ and Fig. 1a shows the arrangement of the *mod(mdg4)* region in the sequence (accession number AE003734), whose principal transcript is 2.2 kilobases (kb) long. Exons I to IV of the 2.2-kb RNA are transcribed from nucleotide positions 153,934 to 151,614; exons V and VI are transcribed from nucleotide positions 136,942 to 137,841 by using the complementary strand as template (Fig. 1a).

To verify the arrangement of sequence AE003734, we used the polymerase chain reaction to amplify DNA from two unrelated wild-type fly strains and an array of nested primers located in coding and non-coding sequences. The amplified DNA fragments for all combinations of primers and for the two strains of fly were of the expected sizes (results not shown). We conclude that the 2.2-kb RNA appears to be encoded by two separate sequences of the gene transcribed in opposite orientations.

We confirmed that the 2.2-kb mRNA was present in fly cells by using a ribonuclease (RNase) protection assay and an RNA probe spanning exons IV and V (Fig. 1b), which revealed a fragment of 598 nucleotides, the size expected in the event of protection by a 2.2-kb mRNA. Other fragments predicted to result from protection by additional *mod(mdg4)* RNAs con-



taining exons IV (435 nucleotides) and V (168 nucleotides) were also detected.

Different mechanisms could account for this 2.2-kb transcript. A switch to a different template by the transcribing enzyme RNA polymerase II, or a programmed genomic rearrangement to bring all the coding sequences to the same DNA strand, could explain why the precursor RNA should contain all six exons. To our knowledge, template switching by RNA polymerase II has never been described, and we have ruled out a genomic rearrangement by using Southern-blot analysis of genomic DNA (results not shown). However, recombination between precursor RNA molecules or *trans* splicing (the use of a splicing donor site from one RNA by the acceptor site from a second⁶) are two other mechanisms that would involve two independent RNA precursors. Of these possible mechanisms, only *trans* splicing has so far been reported in eukaryotes⁷.

The genomic organization of *mod(mdg4)* suggests that two RNA precursors are transcribed to yield the 2.2-kb mRNA. The RNA encoding the *doom* protein (Fig. 1a) is a candidate for one of these precursors⁸. It contains exons I to IV from the 2.2-kb mRNA and extends to nucleotide 131,723 ('class I' precursor in Fig. 1c). A second RNA precursor ('class II' precursor) containing exons V and VI probably originates at nucleotide 136,626 in a promoter region predicted by GENSCAN⁹ (Fig. 1a, c).

Figure 1c shows a simplified model of *trans* splicing and *cis* splicing (donor and acceptor sites on the same molecule) of

Figure 1 Organization of the *mod(mdg4)* locus. **a**, Top, genomic structure. Red and green boxes represent exons in different DNA strands. Arrows indicate the origin of transcription on each strand. A mutant, *mod(mdg4)^{fl}*, is shown here that contains an insertion of the *stalker* retrotransposon at position 137,206 (according to sequence AE003734). Bottom, alternative mRNAs produced by the *mod(mdg4)* locus. Lines connecting exons indicate *cis* splicing; dashed line connects exons transcribed in the opposite orientation. **b**, *Trans*-spliced 2.2-kb mRNA is present in *Drosophila* total RNA. A 598-nucleotide antisense RNA probe containing the last 435 nucleotides from exon IV and the first 163 from exon V of the 2.2-kb *mod(mdg4)* mRNA was synthesized by *in vitro* transcription with T7 polymerase in the presence of [³²P]CTP. The 598-nucleotide riboprobe used in the RNase protection assay is indicated. The 662-nucleotide fragment in lanes 3 and 4 corresponds to the riboprobe plus 64 extra nucleotides transcribed from the plasmid. Lane 1, 50 µg total *Drosophila* RNA; lanes 2 and 3, 50 µg yeast RNA; RNase was added only to lanes 1 and 2. Protected fragments of the predicted sizes are indicated on the left. On the basis of their size, additional bands are probably due to protection of other *mod(mdg4)* RNA species. Lane 4, size markers (nucleotides). **c**, RNA precursors and *cis* splicing or *trans* splicing at the *mod(mdg4)* locus (not to scale). Red and green boxes represent exons transcribed in opposite orientations. Pairing between both precursors facilitates *trans* splicing (circled) and the production of a functional mRNA. Further details are available from the authors.

RNA precursors at the *mod(mdg4)* locus — two nested transcription units in opposite orientation produce two partially complementary RNAs that are able to form double-stranded RNA. It has been shown by using engineered DNAs that sequence complementarity between two RNA molecules can facilitate *trans* splicing in transfected cells¹⁰.

Our results indicate that the evolutionary constraints imposed by a *cis* arrangement of the exons in a gene can be overcome by sequence complementarity between their encoded RNAs. Phylogenetic studies are needed to clarify the evolutionary process that shaped the genomic structure of *mod(mdg4)*. Genome sequences will reveal whether analogous structures are present in other genes and organisms.

Mariano Labrador, Fabien Mongelard, Piedad Plata-Rengifo, Ellen M. Baxter, Victor G. Corces, Tatiana I. Gerasimova

Department of Biology, The Johns Hopkins University, 3400 N Charles Street, Baltimore, Maryland 21218, USA

e-mail: corces@jh.u.edu

1. Burge, C. B., Tuschl, T. & Sharp, P. A. in *The RNA World* 2nd edn (eds Gesteland, R. F., Cech, T. R. & Atkins, J. F.) 525–560 (Cold Spring Harbor Laboratory Press, New York, 1999).
2. Gerasimova, T. I., Gdula, D. A., Gerasimov, D. V., Simonova, O. & Corces, V. G. *Cell* **82**, 587–597 (1995).
3. Gerasimova, T. I. & Corces, V. G. *Cell* **92**, 511–521 (1998).
4. Buchner, K. *et al. Genetics* **155**, 141–157 (2000).
5. Adams, M. D. *et al. Science* **287**, 2185–2195 (2000).
6. Agabian, N. *Cell* **61**, 1157–1160 (1990).
7. Caudevilla, C. *et al. Proc. Natl Acad. Sci. USA* **95**, 12185–12190 (1998).
8. Harvey, A. J., Bidwai, A. P. & Miller, L. K. *Mol. Cell. Biol.* **17**, 2835–2843 (1997).
9. Burge, C. B. & Karlin, S. *Curr. Opin. Struct. Biol.* **8**, 346–354 (1998).
10. Solnick, D. *Cell* **42**, 157–164 (1985).

The impact of specialized enemies on the dimensionality of host dynamics

Ottar N. Bjørnstad^{*†}, Steven M. Sait[‡], Nils C. Stenseth[§], David J. Thompson[‡] & Michael Begon[‡]

^{*} National Center for Ecological Analysis and Synthesis, 735 State Street, Suite 300, Santa Barbara, California 93101-3351, USA

[‡] Population and Evolutionary Biology Research Group, School of Biological Sciences, University of Liverpool, PO Box 147, Liverpool L69 3BX, UK

[§] Division of Zoology, Department of Biology, University of Oslo, PO Box 1050 Blindern, N-0316 Oslo, Norway

Although individual species persist within a web of interactions with other species, data are usually gathered only from the focal species itself. We ask whether evidence of a species' interactions be detected and understood from patterns in the dynamics of that species alone. Theory predicts that strong coupling between a prey and a specialist predator/parasite should lead to an increase in the dimensionality of the prey's dynamics, whereas weak coupling should not. Here we describe a rare test of this prediction. Two natural enemies were added separately to replicate populations of a moth. For biological reasons that we identify here, the prediction of increased dimensionality was confirmed when a parasitoid wasp was added (although this increase had subtleties not previously appreciated), but the prediction failed for an added virus. Thus, an imprint of the interactions may be discerned within time-series data from component species of a system.

Interactions between various natural enemies and their insect prey or hosts represent some of the tightest links in ecological systems^{1,2}. In particular, predominantly specialist enemies may induce coupled host–enemy dynamics³: the abundance of the host species is affected by the abundance of the enemy, which in turn feeds back on the abundance of the host. This dynamic coupling may greatly alter an ecological system's persistence and dynamics^{1–3}. The feedback loop in coupled systems, moreover, generates a delay or lag in the regulation of the host^{4,5}. Thus, theory predicts that an added specialist enemy should result in an increase in the number of lags necessary to explain host dynamics when host dynamics are expressed as a function of host densities at various times in the past^{4–6}. The number of lags or the 'order of density dependence' (the 'embedding dimension' of the systems' dynamics) reflects the number of functionally different interacting groups^{6–8}. Because different life stages of a single species may be functionally distinct, and interactions between different species may be weak, the order need not equate to the number of species present. Thus, as the number of species in a system increases, the 'ecological dimension' (the number of functional groups capable of interacting) increases necessarily, but the dimension of the dynamics need not.

Here we report a direct empirical examination of this relationship between the ecological dimension and the dynamical dimension (the order of density dependence). We added two different natural enemies to laboratory populations of the Indian meal moth (*Plodia interpunctella*): the *P. interpunctella* granulovirus (PiGV) and the parasitoid wasp, *Venturia canescens*. The dimensionality of the dynamics was estimated by nonlinear time-series analysis. We also combine statistical and mathematical modelling to investigate how coupling (or its absence) can be detected in individual time series of the component species of a system, and explore the nature of any changes in dimensionality that may occur.

The study system: basic dynamics

Laboratory cultures of *P. interpunctella* have been studied both experimentally^{2,9–11} and theoretically^{12–15}. Replicated host, host–pathogen and host–parasitoid populations were maintained for about two years (Fig. 1). The parasitoid adults typically lay their eggs

singly in third to fifth instar host larvae^{16,17}, within which they develop and finally kill the host. Parasitoid larvae synchronize their development with that of their host, and adults emerge 21–25 days after parasitization¹⁸. PiGV is a baculovirus that is transmitted through host larval ingestion of virus when either infected cadavers or contaminated medium are eaten^{19,20}. Resistance to infection increases with age: the fifth (final) instar is effectively invulnerable. Overt disease leads to host death, but sublethal effects include reduced fecundity and prolonged development^{9–11,15}.

In host-alone populations, dead adult moths, pupae and larvae

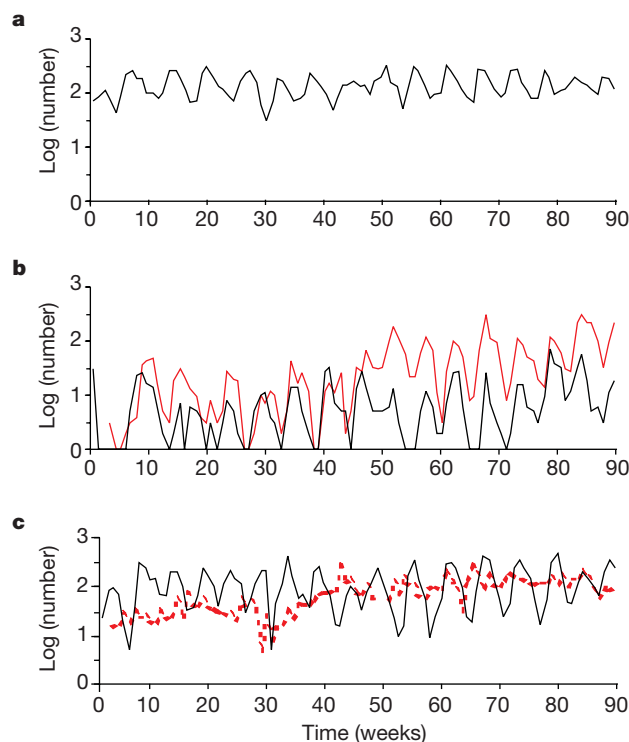


Figure 1 Abundances of the host (thin black line). **a**, Alone; **b**, in the presence of the *Venturia canescens* parasitoid (solid red line); **c**, in the presence of the *Plodia interpunctella* granulovirus (PiGV; dotted red line). The series show representative replicates of each treatment for the first 90 weeks of the experiment.

[†] Present address: Department of Entomology, 501 ASI building, Penn State, University Park, Pennsylvania 16802, USA.

were counted weekly. In host–parasitoid populations, established by adding two parasitoids 14 days after the initial host population had mated, adult wasps were also counted. In host–virus populations, PiGV-infected larvae were added at the same time as the host, and both uninfected and infected larvae (the unit of observation for the virus) were thereafter counted^{2,9–11,16,17}. All time series of the host have dominant periodic behaviour in the 6–7-week range (Table 1). The mean and variability (as measured by the coefficient of variation) of the host in the host–virus system are comparable to those of the host when alone, although the range and amplitude of the cycles are slightly greater (Fig. 1). The host–parasitoid system, in contrast, exhibits greatly depressed host abundance and more violent fluctuations. Furthermore, the period of the fluctuations is slightly shorter (Table 1).

Coupling strengths and dynamic dimensions

The dimension (and order of the density dependence) of each time series was estimated using the nonparametric method of Tong and co-workers^{21–23}, modified to avoid any spurious effects of serial correlations between nearby points in the time series (see Box 1). As determined in previous work¹⁵, the most parsimonious time-series model of the host-alone dynamics is of dimension three (Fig. 2); and, contrary to expectation for coupled enemy–host interactions, there is no evidence of the dimension increasing in the presence of the virus. For the host–parasitoid replicates, however, the dimensions of both the host and the parasitoid series are five—which is conspicuously higher than that of the host when alone or with the virus (Fig. 2; see Box 1). Five gives a significantly better fit than three for these systems, whereas three is significantly superior to five in the host-alone and the host–virus systems (Fig. 2).

The contrast in the dimension of the dynamics is explicable in terms of the different coupling strengths of the virus and parasitoid to the host. To show this, we estimate the nature of the coupling between the different species through nonparametric regression for abundance data^{6,15,24,25} (Fig. 3) using general forms of conventionally accepted functions for all interactions (see Methods). The natural enemy developmental periods are around 3 weeks for both the *Plodia*–*Venturia*¹⁸ and the *Plodia*–PiGV systems^{9–11}. The functions were estimated (see Methods) using a two-dimensional smoothing spline with four degrees of freedom (to avoid making *a priori* assumptions about the functional forms for which we wish to test) and a negative binomial error²⁵. These regression analyses reveal that the number of adult parasitoids emerging is an increasing function of both the numbers of susceptible larvae and adult parasitoids

present 3 weeks earlier (Fig. 3a). Whereas the number of adult hosts decreases rapidly with previous parasitoid abundance, the number increases with host larval abundances (Fig. 3b). This paired set of properties establishes that the parasitoid–host system is fully coupled.

Whereas the number of PiGV-infected larvae is an increasing function of both the number of susceptible larvae and the infected larvae present 3 weeks earlier (Fig. 3c), the number of adult hosts does not decrease with the abundance of previously infected host larvae (Fig. 3d). Thus, the host–virus system is not fully coupled. Enhanced host abundance leads to increased pathogen incidence, but this increase does not feed back negatively on the host. This lack of host–virus coupling is surprising because the virus is a highly specialized enemy that induces significant mortality in the early larval instars of the host^{9–11}. (The alternative hypothesis of a strong virus effect on the host that is hidden because of fast virus dynamics (within a host generation) was considered, but rejected because the virus-infected cadavers (not the virus particles) are the functional unit for transmission^{19,20}.)

The explanation for the unaltered host dimension lies in the strong competition between large larvae¹⁵. In contrast to the parasitoid, which attacks older larvae, the virus is most infectious to younger larvae, and so the virus-induced mortality can be partially compensated for through the life cycle of the host. This is not to say that the virus fails to affect the host dynamics. Indeed, previous studies show that it alters host demography to exaggerate the fluctuations¹⁵. Nevertheless, this virus fails to engage in coupled enemy–host dynamics.

Stage-structure and significant lags

To interpret further the dimension of the dynamics, we develop a simple stage-structured scheme of the interactions¹⁵ (see Box 2). This scheme predicts a three-dimensional model for the host alone (with lags in regulation of 1 week, 3 weeks and 6 weeks), but a five-dimensional model for the host–parasitoid interaction (with additional lags at 4–5 weeks and 7–8 weeks). The three lags in the host-alone model reflect (1) a density-dependent (and negative) interaction at a lag of around 1 week, generated by competition and

Table 1 Summary of the experimental time series

Replicate	Species	n	Mean	Coefficient of variation (C _{95%})	Period
PV1	<i>P. interpunctella</i>	130	12.1	229% (189,262)	40.5
	<i>V. canescens</i>	127	43.9	147% (122,168)	–*
PV2	<i>P. interpunctella</i>	130	8.2	140% (117,162)	42.8
	<i>V. canescens</i>	127	52.2	119% (102,136)	42.8
PV3	<i>P. interpunctella</i>	77	6.2	163% (129,199)	43.5
	<i>V. canescens</i>	74	24.1	126% (105,150)	43.5
PP1	<i>P. interpunctella</i>	74	155.5	79% (67, 89)	47.2
	PiGV	73	69.5	100% (84,114)	47.2
PP2	<i>P. interpunctella</i>	88	151.9	77% (62, 90)	43.3
	PiGV	87	73.4	68% (53, 78)	43.3
PP3	<i>P. interpunctella</i>	91	137.0	81% (70, 92)	45.8
	PiGV	90	85.2	67% (57, 77)	–
P1	<i>P. interpunctella</i>	95	105.6	62% (52, 70)	44.3
P2	<i>P. interpunctella</i>	48	108.2	73% (59, 87)	52.5
P3	<i>P. interpunctella</i>	108	132.4	59% (52, 66)	45.0

PV1–PV3 signify the three host–parasitoid replicates (*P. interpunctella* is the host, *V. canescens* is the parasitoid). PP1–PP3 signify the three host–virus replicates (*P. interpunctella* is the host, PiGV is the virus infected larvae). P1–P3 signify the three host-alone replicates. n, total length of the time series; mean, average abundance; coefficient of variance denotes 100s.d./mean, where s.d. is the standard deviation. 95% Confidence intervals (C_{95%}) are calculated by bootstrapping the series (with 10,000 resamples). Period denotes the dominant period in the periodogram (in weeks). * Subdominant period: 38.5 days.

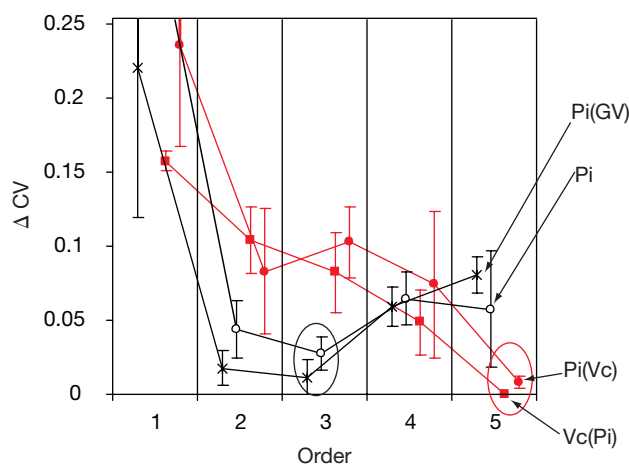


Figure 2 Cross-validation for the order of density dependence according to the method of Tong and co-workers^{21–23} (Box 1). Cross-validation error is plotted against order for each of the three treatments. Shown are the cross-validation profiles for *P. interpunctella* (Pi), the host alone; Pi(Vc), the host in the presence of the parasitoid; Pi(GV), the host in the presence of the virus; Vc(Pi), the parasitoid in the presence of the host. For each replicate, the cross-validation error was calculated relative to the optimal order. ΔCV represents the average loss in predictive ability across all replicates. The error bars represent one standard error. The parsimonious orders (circles) are three for the host alone and the host in the presence of the virus, but five for the host–parasitoid interaction. See Supplementary Information for details.

cannibalism of large larvae on small; (2) a second density-dependent interaction at a lag of around 3 weeks, generated by competition amongst small larvae and cannibalism of large larvae on eggs; and (3) a positive term at a lag of around one host generation (~ 6 weeks), representing host reproduction (but discounted by intra-stage competition).

The host–parasitoid model predicts the dimension to increase by two, but the delayed density-dependent structure contributes a total of five new terms to the host dynamics (Box 2). Three of these take the same form: the dependences predicted by the host-alone model, discounted by the rate (μ_p) at which parasitoid attack rate (in the vicinity of the equilibrium) increases with increasing parasitoid abundance. Each of these terms appears about 2 weeks later in the new model as compared with the original term. Two weeks is the approximate delay between the peak period of parasitoid attacks, on the largest host larvae, and the time when unparasitized hosts would have emerged as adults^{9–11}). Thus, we see—three times—the loss of

hosts to parasitoids leading to a dilution of the interactions in the host-alone model. There is less reproduction, less competition and less cannibalism because there are fewer hosts, as a result of their having been attacked by adult parasitoids present ~ 2 weeks previously.

The fourth addition further discounts the host reproduction term (lag 6). Thus, discounting because of losses of larvae is now attributable not only to intra-stage competition (as in the host-alone model), but also because increased abundance of larvae leads to increased parasitoid attack rates.

Finally, the fifth term, operating with a 1-week lag, appears to describe the effect of survival against parasitoid attack, as it enters in the form $A_t = \mu_p A_{t-1}$, where A_t represents the number of adult moths at time t (Box 2), that is, μ_p here is a survival rate. Moreover, because μ_p is low when parasitoids experience intense ‘interference’ at equilibrium but is high when they do not, the term appears also to capture, as predicted by theory²⁶, the potentially damping effects

Box 1

Time-series estimates of the order

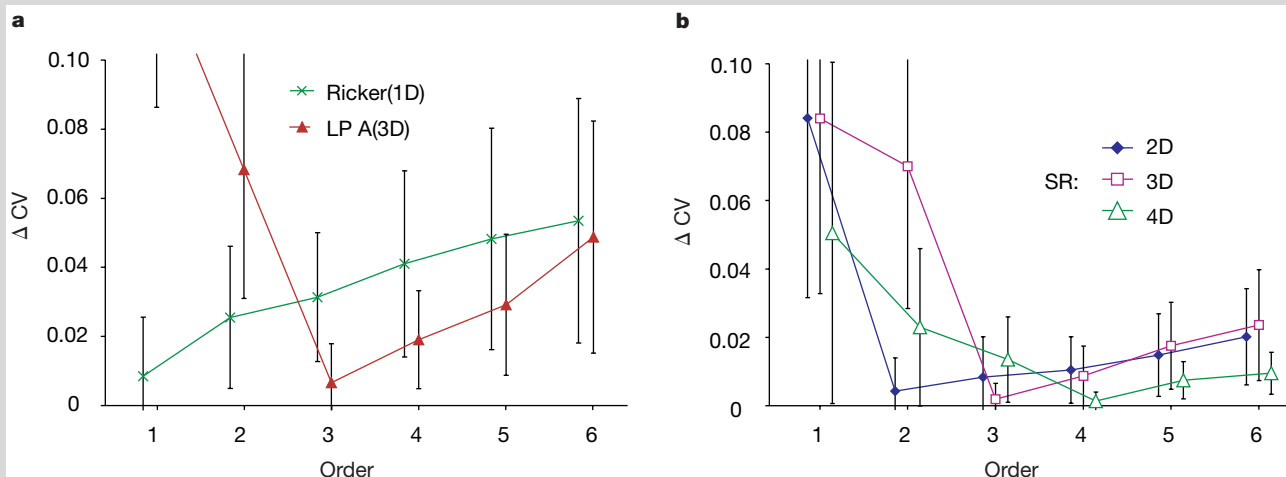
Royma⁴ developed the theory for the interrelation between the dynamical dimension and the order of density dependence of an ecological time series. This order may therefore be studied when trying to understand the dynamics of a system^{5,8}. The classical estimation of the order of a time series relies on the goodness of fit of linear models of various complexities³⁵. However, several alternative methods have been proposed in order not to constrain analyses to be valid only for roughly linear dynamical systems. We estimate the order of each time series according to the order-consistent, nonparametric method of Tong and co-workers^{21–23}. This method does not require assumptions about linearity (or any other functional form, except the assumption of smoothness³⁶) and uses cross-validation of the locally linear regression against lagged abundances to estimate the expected future abundance on the basis of a population’s trajectory. The candidate nonparametric model that presupposes the correct order will be the best at forecasting, and will thus minimize the cross-validation error²¹. As there is fairly strong correlation between adjacent observations and this may induce spurious results³⁷, we remove a whole generation (7 weeks, that is, 3 weeks at each side of the target week) during the cross-validation (see Methods). Pooling across replicates, dimension 5 was a significantly better fit to the data than dimensions 1–4 for the time series of host with parasitoid (and parasitoid with host) (Fig. 2), whereas for host alone and host with virus, dimension 3 was the best fit—significantly better than dimensions 1, 4 and 5 for the host–virus systems, and than dimensions 1 and 4 for the host alone.

Testing the method

The order-consistent nonparametric method is known to provide correct estimates of the order of long time series^{21–23}. Whether it is possible to

identify high orders correctly from the type of data obtained in ecological studies, however, is unknown. We therefore undertook an analysis of the power of the method when applied to time series from three well known age-/stage-structured ecological models representing many dimensional complexities. In each instance, we simulate the models for 150 time steps and discard the first 50 observations. The resulting time series—of comparable length (100 observations) to the replicates analysed above—are subjected to identical analyses to those of the real data. The simplest model is a (non-age-structured) one-dimensional stochastic Ricker model³⁸. The second model is the three-dimensional LPA-model for the stage-structured dynamics of the flour beetle³⁰. The third model is the SR model for fish stock/recruitment dynamics³¹, in which the dimension can be anywhere from two to four depending on the number of age classes. (See Methods for details of the models and parameters used.)

The figure shows the cross-validation (CV) profiles across 25 realizations of each model. For the Ricker model (a), the order is correctly identified as 1 in most cases. Note that the bars represent standard deviations and not standard errors. Furthermore, the order is correctly identified as 3 in most cases of the LPA model (see figure), and as 2, 3, and 4, respectively, in the SR model (b). In the Supplementary Information we detail results for a wide range of model parameters for each of the three sample models. The overall conclusions of the simulation are that the nonparametric method can correctly identify the order (of at least up to four) of time series from diverse ecological systems on the basis of 100 observations. The order tends to be biased upwards for highly nonlinear versions of the models. The order is at most increased by one, however, and the fit of the larger model is only very marginally better (by ΔCV around 0.01)



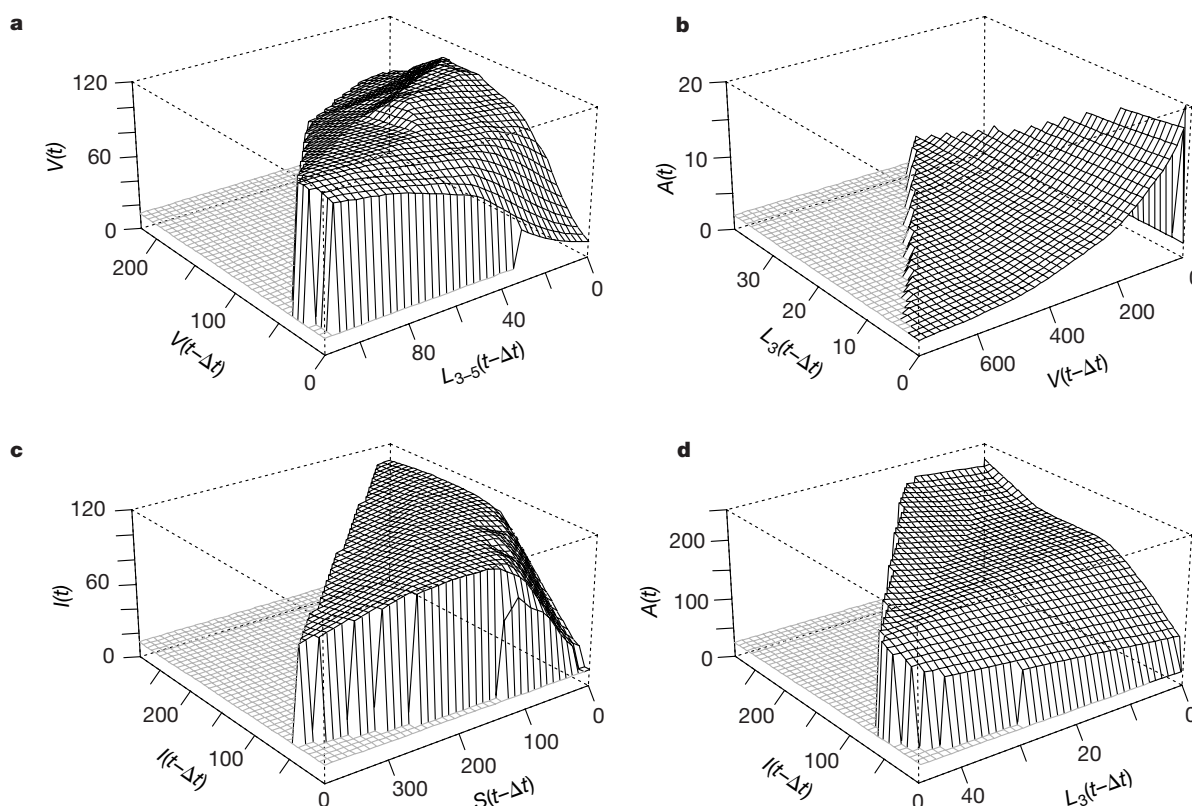


Figure 3 The estimated functions. **a**, Abundance of adult parasitoids, V , as a function of the number of parasitoids and susceptible host larvae (instar 3–5), L_{3-5} , 3 weeks previously (function H in Methods). **b**, Abundance of adult hosts, A , versus abundance of host (instar 3) larvae, L_3 , and the three parasitoid cohorts that may attack them 3 weeks previously (function K in Methods). **c**, Abundance of granulovirus-infected larvae, I , versus the number of infected larvae and susceptible larvae (instar 1–4), S , 3 weeks

previously (function F in Methods). **d**, Abundance of adult hosts, A , versus abundance of host (instar 3) larvae, L_3 , and infected larvae 3 weeks previously (function G in Methods). We estimate the functions by using a two-dimensional smoothing spline. All functions are statistically significant at a nominal 1% level (see Methods). The grey-shaded grids along the xy -planes signify states for which no data are available and to which no functions are fitted.

on dynamics of parasitoid attack acting without a delay: more parasitoid interference leads to more immediate (as opposed to delayed) feedback on host dynamics.

Thus, the move from three to five dimensions in the presence of the parasitoid turns out not simply to be a case of the original hosts' three dimensions being supplemented by two new host–parasitoid dimensions. Instead, more subtly, there appears to be a combination of direct, trophic effects (entirely new terms) and indirect, modulating effects (modifications of existing terms, sometimes with a time delay). The new terms sometimes increase the dimensionality but at other times do not.

Furthermore, there is compelling evidence within the time-series data themselves for the precise lag structures predicted by the stage-structured models. Using Wald tests²⁷ (see Methods), highly significant lags are detected in the host data from the host–parasitoid series at lags of 1, 3, 4–5, 6 and 7–8 weeks. In contrast, host series when alone or in the presence of the virus exhibit highly significant lags at 1, 3 and 6 weeks, but no evidence of lagged regulation at 4–5 or 7–8 weeks. This striking congruence between data and models reinforces our confidence in the results of both the mechanistic predictions and the statistical analyses.

Conclusions

Our results show that specialist enemies can, as theory predicts, increase the dimension of host dynamics through complete coupling (*Venturia*–*Plodia*), but also that the increase in dimensional complexity can be counterbalanced if coupling is weak (PiGV–*Plodia*). There may be a direct connection in ecological systems between the number of identifiable interacting groups and the

dynamic dimension—but such a connection is not inevitable. This helps explain why certain keystone species embedded in rich ecological communities exhibit low-dimensional dynamics⁶. Potentially, it may also illuminate the enigmatic nature of biological control²⁸. *Plodia* is a serious stored-food pest²⁹, and pathogens (such as PiGV) and parasitoids (such as *Venturia*) are prime candidates for its biological control. Although they are both specialist enemies, one (the parasitoid) serves to depress host densities (Table 1) and the other (the virus) is ineffective. The difference in efficiency is related to the way the parasitoid engages in coupled interactions with the host but the virus does not.

Furthermore, our findings show how even when specialist enemies (or other additional species) increase the dynamic dimension of an ecological system, there may be a complex, but tractable, link between the number of added interactants and the number of added dimensions. Finally, we are excited to show that, with appropriate analysis, the imprint of the detailed interactions in a system's dynamics can be discerned from time-series data of a component species. □

Methods

The locally linear regression

A gaussian product kernel was used. The bandwidth of the kernel was also optimized using cross-validation. We carried out the analyses on time series for which the first 20 observations (~3 host generations) were discarded to allow for a transient period. The data were square-root transformed to stabilize the variance. Overall results are given in Fig. 2 and Box 1; for the results for individual replicates, see Supplementary Information.

Ecological models in Box 1

The 'Ricker model' is a one-dimensional model, $N_{t+1} = N_t \exp[r(1 - N_t)]u_t$, where N is

A predictive model for dimensional changes

To predict the impact of the parasitoid on the host dynamics, we first develop a model for the host dynamics. We then add parasitoid–host interaction to the model to calculate its effects. We consider the following five stages in the host life-cycle: E, eggs; S, small larvae (consisting of instars 1–3); L, large larvae (consisting of instars 4–5); P, pupae; A, adults. The developmental times between each of these are in the order of (but on average a little longer than) a week (range from 5 to 10 days; see details in ref. 15). Because adults do not require any resources, the number of eggs produced at time t is directly proportional to the number of adults, $E_t = rA_t$, where r is the *per capita* reproduction. For simplicity, we assume the large larvae to be the most important cannibals of eggs and larvae. Thus, survival of eggs may decrease due to density-dependent cannibalism by these. In this way, $S_t = E_{t-1}\Gamma(L_{t-1})$, where Γ is a decreasing function of L_{t-1} . Within-stage competition and cannibalism by larger larvae reduce the survival of small larvae to the next stage: $L_t = S_{t-1}\Lambda(S_{t-1}, L_{t-1})$, where Λ is a decreasing function of both S_{t-1} and L_{t-1} . Assuming competitive dominance of the large larvae, the survival of these depends only on within stage competition. In this way the number of pupae is:

$$P_t = L_{t-1}\Omega(L_{t-1}) \quad (1)$$

where Ω is decreasing in L_{t-1} . For simplicity we assume density-independent adult emergence (at rate c) from the pupae:

$$A_t = cP_{t-1} \quad (2)$$

We can use a Taylor expansion on a log-scale on this set of five life-cycle equations to predict the dimension of the dynamics. We denote intra-stage density dependence within the small larval stage and large larval stage (in the vicinity of the stable or unstable equilibrium) by β_s and β_l , respectively, and the influence of large larvae on eggs and small larvae by γ_e and γ_s , respectively. We can then write a model in delay coordinates as:

$$\begin{aligned} \log A_t = & -\log A_{t-1}[\gamma_s] - \log A_{t-3}[(1 - \beta_s)\gamma_e] \\ & + \log A_{t-6}[rcP^*(1 - \beta_s)(1 - \beta_l)] + \text{const} \end{aligned} \quad (3)$$

where the asterisk denotes evaluations at the equilibrium, and const represents the collection of constant terms. Equation (3) predicts that life-

cycle interactions within the host result in three-dimensional dynamics with lags in regulation at week 1, week 3 and week 6. The exact delays result from considering the duration of each life-stage¹⁵. The first and second lag reflect cannibalism of small larvae and eggs, respectively. But note that the latter is discounted by within-stage competition. The third lag reflects reproduction (but as modified by intra-stage competition in small and large larvae).

The parasitoid–host interaction

In the presence of the parasitoid we need to modify equations (1) and (2) to take into account that a proportion of the large larvae will become adult parasitoids (see Methods): $V_t = kL_{t-2}\Omega(L_{t-2})h(V_{t-2}, L_{t-2})$. The complementary proportion will escape parasitism and metamorphose into adult hosts $A_t = cL_{t-2}\Omega(L_{t-2})[1 - h(V_{t-2}, L_{t-2})]$. We can use a Taylor expansion (on a log-scale) on this modified set of life-cycle equations to predict the changes to host dimensional complexity when interacting with a parasitoid. We let μ_l and μ_p denote, respectively, how increases in large *Plodia* larvae and increases in adult parasitoid abundance affect the parasitoid attack rate (in the vicinity of the stable or unstable equilibrium). All other symbols are as defined above. The resultant delay coordinate representation is for *Plodia*–*Venturia*:

$$\begin{aligned} \log A_t = & -\log A_{t-1}\{\gamma_s - \mu_p\} \\ & - \log A_{t-3}\{(1 - \beta_s)\gamma_e + \gamma_s\mu_p\} \\ & - \log A_{t-4/5}\{(1 - \beta_s)\gamma_e\mu_p\} \\ & + \log A_{t-6}\{rcP^*(1 - \beta_s)(1 - \beta_l) - rcP^*(1 - \beta_s)\mu_l\} \\ & - \log A_{t-7/8}\{rcP^*(1 - \beta_s)(1 - \beta_l)\mu_p\} \\ & + \text{const} \end{aligned} \quad (4)$$

Equation (4) is five-dimensional, with the same three lags as equation (3) but with additional lags at 4–5 weeks and 7–8 weeks. A total of five additional terms are induced. In the equation, these are placed in a separate column on the right to aid readability. The model thus predicts an increased dimensionality when the parasitoid was added, though the nature of the increase has several subtleties.

the scaled density, r ($= 1.5$) is the growth rate, and u is a sequence of log-normal variates with unit mean. In the simulations, we assumed a coefficient of variation of 10%. The ‘LPA model’ is the three-dimensional larvae–pupae–adult (LPA) model for the dynamics of *Tribolium* flour beetle populations³⁰: $L_{t+1} = bA_t \exp[-c_d L_t - c_{ca} A_t + E_{1t}]$; $P_{t+1} = L_t(1 - \mu_l)$; $A_{t+1} = P_t \exp[-c_{pa} A_t] + A_t(1 - \mu_a)$, where b ($= 6.6$) represents reproduction; c_{cl} ($= 1.2 \times 10^{-2}$), c_{ca} ($= 1.2 \times 10^{-2}$) are cannibalism rates of eggs by pupae and larvae, respectively; c_{pa} ($= 0.05$) is adult cannibalism on pupae; and μ_l ($= 0.2$) and μ_a ($= 0.8$) are larval and adult mortality rates. E_{1t} is a sequence of normal variates with mean zero and standard deviation 0.35.

The ‘SR model’ is the stock-recruitment model of fish dynamics that encompasses the interactions between young of the year (X), juveniles (Y) and mature individuals (S)³¹: $X_t = bS_t\alpha_t$; $Y_{t+1} = X_t \exp[-\beta \ln(X_t) - \gamma \ln(Y_t)]$; $S_{t+1} = (1 - \mu_y)Y_t + (1 - \mu_s)S_t$, where b ($= 13$) is the reproduction rate, and α is a sequence of log-normal variates (with a mean of 1 and a coefficient of variation 0.3); β and γ are juvenile competition and cannibalism rates; μ_y ($= 0.1$) and μ_s are juvenile and mature mortality rates. The dimension of the model depends on the mortality rates of the mature individuals, ranging from 2 (when $\mu_s = 1$) to 3 or 4 ($\mu_s < 1$). In the simulation we use $\beta = 0.5$ and $\gamma = 0.4$ when $\mu_s = 1$, and $\beta = 0.3$ and $\gamma = 0.7$ when $\mu_s < 1$. (In the Supplementary Information we consider a wide range of parameter values for each of the three models.)

Virus–host and parasitoid–host coupling

Transmission of the virus is through contact between previously infected larvae, I , and susceptible larvae, S , taken to consist of the first four larval instars. The proportion, $f()$, of susceptible larvae that becomes newly infective in the period from $t - \Delta t$ to t is a function of the abundance of infectives and susceptibles: $I_t = S_{t-\Delta t}f(I_{t-\Delta t}, S_{t-\Delta t}) = F(I_{t-\Delta t}, S_{t-\Delta t})$, where Δt is the time taken for disease development. Full virus–host coupling implies that as virus abundance increases with S , host abundance decreases with I . For strictly lethal infections it should do so. But infection may not necessarily be lethal^{2–11,32}, so the number of adult *Plodia*, A , can be modelled as $A_t = m_s S_{t-\Delta t}[1 - f()] + em_s S_{t-\Delta t}f() = G(I_{t-\Delta t}, S_{t-\Delta t})$, where m_s is the fraction of uninfected larvae that successfully metamorphose into adults. The subscript s indicates that this fraction may be density dependent. The

parameter $(1 - e)$ measures virus-induced mortality; $m_s e$ is thus the fraction of infected larvae that successfully metamorphose into adults.

‘Transmission’ of the parasitoid occurs through adult parasitoids (the ‘infective stage’) ovipositing into susceptible larvae. The number of susceptible larvae, L , consist of larval instars 3–5 (refs 16–18). The number of new adult parasitoids, V , is a function of the abundance of adult parasitoids and susceptible host larvae one development period previously: $V_t = kL_{t-\Delta t}h(V_{t-\Delta t}, L_{t-\Delta t}) = kH(V_{t-\Delta t}, L_{t-\Delta t})$, where Δt is the time taken for parasitoid development, and k is the hatching rate of adult parasitoids from parasitized larvae. The function $h()$ governs the proportion of larvae that become parasitized. Classically $h() = 1 - \exp(-mV_{t-\Delta t})$ ³³, so that $H()$ is typically an increasing function in V and L . Parasitization is usually lethal¹⁸, so the number of adult *Plodia* emerging will be a function of the unparasitized larvae: $A_t = m'_s L_{t-\Delta t}[1 - h()] = K(V_{t-\Delta t}, S_{t-\Delta t})$. The parameter m'_s is the fraction of unparasitized larvae that successfully metamorphose into adults. The subscript s indicates that this fraction may be density dependent.

When we estimated the functions nonparametrically we used two-dimensional smoothing splines with four degrees of freedom^{15,25}, and a negative binomial error model^{25,34}. In Fig. 3, the information is pooled across the three replicates of each treatment¹⁵. Zero observations in the response (that is no adults observed) were omitted from the analyses because these induced singularities in the estimation. All functions are statistically significant (Fig. 3a: $F_{280,4} = 21.5$, $P < 0.001$; 3b: $F_{286,4} = 5.55$, $P < 0.001$; 3c: $F_{283,4} = 19.1$, $P < 0.001$; 3d: $F_{253,4} = 5.54$, $P < 0.001$). The shape parameter, κ , in the negative binomial is around unity or lower for the *Plodia*–*Venturia* system (*Plodia* $\kappa = 0.69$, s.e. = 0.07; *Venturia* $\kappa = 1.06$, s.e. = 0.10), and slightly higher than that in the *Plodia*–PiGV system (*Plodia* $\kappa = 1.62$, s.e. = 0.13; PiGV $\kappa = 3.08$, s.e. = 0.26). The results for the individual replicates are near identical to the treatment-wise analysis, although the power is somewhat lower (see Supplementary Information for details).

Testing for significant lags

We use the Wald test²⁷ to test for significant lags in regulation. A generalized linear model was constructed relating log-abundance to log-abundance at lags: 1, 3, 6, 4–5, and 7–8 for the time series of the host alone and in the presence of the two specialist enemies (Box 2). A

log-link was used and the error was assumed to follow an overdispersed Poisson process³⁴. Lags falling between sampling intervals were studied by averaging across the two adjacent sampling intervals (that is, lag 4–5 is taken as average of abundance in the week 4 and week 5 previously). A constant of unity was added to the lagged variables before log-transformation. The results (summarized across the replicates¹⁵) are as follows: *Plodia*–*Venturia* series: (lag 1: $t = 13.48$, $P < 0.01$; lag 3: $t = -3.03$, $P < 0.01$; lag 4–5: $t = 3.38$, $P < 0.01$; lag 6: $t = 3.37$, $P < 0.01$; lag 7–8: $t = -4.35$, $P < 0.01$); *Plodia*–PiGV series: (lag 1: $t = 2.30$, $P < 0.02$; lag 3: $t = -4.62$, $P < 0.01$; lag 4–5: $t = -0.95$, $P = 0.34$; lag 6–7: $t = 5.58$, $P < 0.01$; lag 7–8: $t = -0.62$, $P = 0.53$); *Plodia* alone series: (lag 1: $t = 3.86$, $P < 0.01$; lag 3: $t = -3.49$, $P < 0.01$; lag 4–5: $t = 0.59$, $P = 0.56$; lag 6: $t = 6.75$, $P < 0.01$; lag 7–8: $t = 0.72$, $P = 0.47$). Note that PiGV increases the generation time of the host by about half a week¹⁵. The significance of regulation is therefore tested at a lag of 6–7 weeks. The results for the individual replicates are near identical to the treatment-wise average, although the power is somewhat lower (see Supplementary Information for details).

Received 12 October; accepted 29 November 2000.

- Hassell, M. P. & May, R. M. Generalist and specialist natural enemies in insect predator–prey interactions. *J. Anim. Ecol.* **55**, 923–940 (1986).
- Begon, M., Sait, S. M. & Thompson, D. J. Predator–prey cycles with period shifts between two- and three-species systems. *Nature* **381**, 311–315 (1996).
- Murdoch, W. W. & Stewart-Oaten, A. Predation and population stability. *Adv. Ecol. Res.* **9**, 1–131 (1975).
- Royama, T. *Analytical Population Dynamics* (Chapman & Hall, London, 1992).
- Turchin, P. Rarity of density dependence or population regulation with lags? *Nature* **344**, 660–663 (1990).
- Stenseth, N. C., Falck, W., Bjørnstad, O. N. & Krebs, C. J. Population regulation in snowshoe hare and lynx populations: asymmetric food web configurations between the snowshoe hare and the lynx. *Proc. Natl Acad. Sci. USA* **94**, 5147–5152 (1997).
- Schaffer, W. M. Ecological abstraction: The consequence of reduced dimensionality in ecological models. *Ecol. Monogr.* **51**, 383–401 (1981).
- Berryman, A. A. *Population Systems: a General Introduction* (Plenum, New York, 1981).
- Sait, S. M., Begon, M. & Thompson, D. J. Long-term population dynamics of the Indian meal moth *Plodia interpunctella* and its granulosis virus. *J. Anim. Ecol.* **63**, 861–870 (1994).
- Sait, S. M., Begon, M. & Thompson, D. J. The influence of larval age on the response of *Plodia interpunctella* to a granulosis virus. *J. Invert. Pathol.* **63**, 107–110 (1994).
- Sait, S. M., Begon, M. & Thompson, D. J. The effect of a sublethal baculovirus infection in the Indian meal moth, *Plodia interpunctella*. *J. Anim. Ecol.* **63**, 541–550 (1994).
- Gurney, W. S. C., Nisbet, R. M. & Lawton, J. H. The systematic formulation of tractable single-species population models incorporating age structure. *J. Anim. Ecol.* **52**, 479–495 (1983).
- Gurney, W. S. C. & Nisbet, R. M. Fluctuation periodicity, generation separation, and the expression of larval competition. *Theor. Popul. Biol.* **28**, 150–180 (1985).
- Briggs, C. J., Sait, S. M., Begon, M., Thompson, D. J. & Godfray, H. C. J. What causes generation cycles in populations of stored product moths? *J. Anim. Ecol.* **69**, 352–366 (2000).
- Bjørnstad, O. N. *et al.* Population dynamics of the Indian meal moth: demographic stochasticity and delayed regulatory mechanisms. *J. Anim. Ecol.* **67**, 110–126 (1998).
- Sait, S. M. *et al.* *Venturia canescens* parasitizing *Plodia interpunctella*: host vulnerability – a matter of degree. *Ecol. Entomol.* **20**, 199–201 (1995).
- Sait, S. M., Begon, M., Thompson, D. J., Harvey, J. A. & Hails, R. S. Factors affecting host selection in an insect host–parasitoid interaction. *Ecol. Entomol.* **22**, 225–230 (1997).
- Harvey, J. A., Harvey, I. F. & Thompson, D. J. Flexible larval growth allows use of a range of host sizes by a parasitoid wasp. *Ecology* **75**, 1420–1428 (1994).

- Boots, M. Cannibalism and the stage-dependent transmission of a viral pathogen of the Indian meal moth, *Plodia interpunctella*. *Ecol. Entomol.* **23**, 118–122 (1998).
- Knell, R. J., Begon, M. & Thompson, D. J. Transmission of *Plodia interpunctella* granulosis virus does not conform to the mass action model. *J. Anim. Ecol.* **67**, 592–599 (1998).
- Cheng, B. & Tong, H. On consistent nonparametric order determination and chaos. *J. R. Statist. Soc. B* **54**, 427–449 (1992).
- Cheng, B. & Tong, H. Orthogonal projection, embedding dimension and sample size in chaotic time series from a statistical perspective. *Phil. Trans. R. Soc. Lond. A* **348**, 325–341 (1994).
- Yao, Q. & Tong, H. Quantifying the influence of initial values on non-linear prediction. *J. R. Statist. Soc. B* **56**, 701–725 (1994).
- Ellner, S. P. *et al.* Noise and nonlinearity in measles epidemics: Combining mechanistic and statistical approaches to population modeling. *Am. Nat.* **151**, 425–440 (1998).
- Bjørnstad, O. N., Fromentin, J.-M., Stenseth, N. C. & Gjøsæter, J. A new test for density-dependent survival: the case of coastal cod populations. *Ecology* **80**, 1278–1288 (1999).
- Free, C. A., Beddington, J. R. & Lawton, J. H. Inadequacy of simple-models of mutual interference for parasitism and predation. *J. Anim. Ecol.* **46**, 543–554 (1977).
- Sokal, R. R. & Rohlf, F. J. in *Biometry* 1 Ch. XIX, 887 (W. H. Freeman, New York, 1995).
- Murdoch, W. W. & Briggs, C. J. Theory of biological control: recent developments. *Ecology* **77**, 2001–2013 (1996).
- Tabashnik, B. E. & McGaughey, W. H. Resistance risk assessment for single and multiple insecticides: Responses of Indian meal moth (Lepidoptera: Pyralidae) to *Bacillus thuringiensis*. *J. Econ. Entomol.* **87**, 835–841 (1994).
- Costantino, R. F., Desharnais, R. A., Cushing, J. M. & Dennis, B. Chaotic dynamics in an insect population. *Science* **275**, 389–391 (1997).
- Bjørnstad, O. N., Fromentin, J.-M., Stenseth, N. C. & Gjøsæter, J. Cycles and trends in cod population. *Proc. Natl Acad. Sci. USA* **96**, 5066–5071 (1999).
- Grenfell, B. T. & Dobson, A. P. Ecology of infectious diseases in natural populations. (Cambridge Univ. Press, Cambridge, 1995).
- Nicholson, A. J. & Bailey, V. A. The balance of animal populations. *Proc. Zool. Soc. Lond.* **3**, 551–598 (1935).
- Venables, W. N. & Ripley, B. D. in *Modern Applied Statistics with S-plus* 1–462 (Springer, New York, 1994).
- Wei, W. W. *Time Series Analysis* (Addison Wesley, California, 1989).
- Green, P. J. & Silverman, B. W. in *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach* Ch. 1–IX, 182 (Chapman & Hall, London, 1994).
- Ellner, S. & Turchin, P. Chaos in a noisy world: New methods and evidence from time-series analysis. *Am. Nat.* **145**, 343–375 (1995).
- Ricker, W. E. Stock and recruitment. *J. Fish. Res. Board Canada* **11**, 559–623 (1954).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

Funding was received from the National Center for Ecological Analysis and Synthesis (O.N.B.) (a Center funded by the NSF, the University of California Santa Barbara and the State of California), from the Norwegian National Science Foundation (O.N.B., N.C.S.) and from NERC (M.B., S.M.S. and D.J.T.). P. Amarasekare, A. Dobson and B. Grenfell commented on the manuscript.

Correspondence and requests for materials should be addressed to O.N.B. (e-mail: onb1@psu.edu).

Massive gene decay in the leprosy bacillus

S. T. Cole*, K. Eiglmeier*, J. Parkhill†, K. D. James†, N. R. Thomson†, P. R. Wheeler‡, N. Honoré*, T. Garnier*, C. Churcher†, D. Harris†, K. Mungall†, D. Basham†, D. Brown†, T. Chillingworth†, R. Connor†, R. M. Davies†, K. Devlin†, S. Duthoy*, T. Feltwell†, A. Fraser†, N. Hamlin†, S. Holroyd†, T. Hornsby†, K. Jagels†, C. Lacroix*, J. Maclean†, S. Moule†, L. Murphy†, K. Oliver†, M. A. Quail†, M.-A. Rajandream†, K. M. Rutherford†, S. Rutter†, K. Seeger†, S. Simon*, M. Simmonds†, J. Skelton†, R. Squares†, S. Squares†, K. Stevens†, K. Taylor†, S. Whitehead†, J. R. Woodward† & B. G. Barrell†

*Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France

†Sanger Centre, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK

‡Veterinary Laboratories Agency, Weybridge, Woodham Lane, New Haw, Addlestone, Surrey KT15 3NB, UK

Leprosy, a chronic human neurological disease, results from infection with the obligate intracellular pathogen *Mycobacterium leprae*, a close relative of the tubercle bacillus. *Mycobacterium leprae* has the longest doubling time of all known bacteria and has thwarted every effort at culture in the laboratory. Comparing the 3.27-megabase (Mb) genome sequence of an armadillo-derived Indian isolate of the leprosy bacillus with that of *Mycobacterium tuberculosis* (4.41 Mb) provides clear explanations for these properties and reveals an extreme case of reductive evolution. Less than half of the genome contains functional genes but pseudogenes, with intact counterparts in *M. tuberculosis*, abound. Genome downsizing and the current mosaic arrangement appear to have resulted from extensive recombination events between dispersed repetitive sequences. Gene deletion and decay have eliminated many important metabolic activities including siderophore production, part of the oxidative and most of the microaerophilic and anaerobic respiratory chains, and numerous catabolic systems and their regulatory circuits.

Leprosy, one of the oldest recorded diseases, remains a major public health problem. Although prevalence has been reduced extensively by WHO multidrug therapy and vaccination with BCG^{1,2}, the incidence of the disease remains worrying with more than 690,000 new cases reported annually³. In 1873, in the first convincing association of a microorganism with a human disease, Armauer Hansen⁴ discovered the leprosy bacillus in skin biopsies but failed to culture *Mycobacterium leprae*. A century later, the nine-banded armadillo⁵ was used as a surrogate host, enabling large quantities of the bacillus to be isolated for biochemical and physiological studies. Subsequent efforts to demonstrate multiplication in synthetic media have been equally fruitless, although metabolic activity can be detected⁶. The exceptionally slow growth of the bacillus, which has a doubling time of ~14 days (ref. 7), may contribute to these failures.

The means of transmission of leprosy is uncertain but, like tuberculosis, the infection is thought to be spread by the respiratory route because lepromatous patients harbour bacilli in their nasal passages. The bacterium accumulates principally in the extremities of the body where it resides within macrophages and infects the Schwann cells of the peripheral nervous system. Lack of myelin production by infected Schwann cells, and their destruction by host-mediated immune reactions, leads to nerve damage, sensory loss and the disfigurement that, sadly, are the hallmarks of leprosy.

Genome features and reductive evolution

Sequence analysis. The complete genome sequence of *M. leprae* contains 3,268,203 base pairs (bp), and has an average G+C content of 57.8% (Table 1). These values are much lower than those reported for the *M. tuberculosis* genome, which comprises ~4,000 genes, 4,411,532 bp and 65.6% G+C (ref. 8). From detailed pairwise comparisons of both genome and proteome sequences^{8,9}, only 49.5% of the *M. leprae* genome contains protein-coding genes, whereas 27% contains recognizable pseudogenes (inactive reading frames with functional counterparts in the tubercle bacillus). The remaining 23.5% of the genome does not appear to be coding, and may correspond to regulatory sequences or even gene remnants

mutated beyond recognition. The distribution of the 1,116 pseudogenes is essentially random (Fig. 1), and if these are excluded 1,604 potentially active genes remain, of which 1,439 are common to both pathogens. Among the 165 genes with no orthologue in *M. tuberculosis* are 29 for which we can attribute functions. Many of the 136 residual coding sequences in *M. leprae*, which show no similarity to known genes, may also represent pseudogenes as they are shorter than average (Table 1) and occur in regions of low gene density (Fig. 1).

Reductive evolution. Assuming that the genome of *M. leprae* was once topologically equivalent and similar in size to those of all other mycobacteria (~4.4 Mb)^{10–12}, then extensive downsizing and rearrangement must have occurred during evolution. If all the genes in the ~3.3 Mb *M. leprae* genome were active, one would expect 3,000 proteins as compared with the 4,000 predicted in *M. tuberculosis*. Comparative proteome analysis detected only 391 soluble protein species¹³, compared with ~1,800 in *M. tuberculosis*¹⁴, indicating that the pseudogenes are translationally inert. Thus, since diverging from the last common mycobacterial ancestor, the leprosy bacillus may have lost more than 2,000 genes.

Reductive evolution is documented in obligate intracellular parasites, such as *Rickettsia* and *Chlamydia* spp., and in some endosymbionts¹⁵, because genes become inactivated once their functions are no longer required in highly specialized niches. This process may have naturally defined the minimal gene set for a

Table 1 Comparison of genome features

Feature	<i>M. leprae</i>	<i>M. tuberculosis</i>
Genome size (bp)	3,268,203	4,411,532
G+C (%)	57.79	65.61
Protein coding (%)	49.5	90.8
Protein-coding genes (no.)	1,604	3,959
Pseudogenes (no.)	1,116	6*
Gene density (bp per gene)	2,037	1,114
Average gene length (bp)	1,011	1,012
Average unknown gene length (bp)	338	653

*Excluding IS elements.

pathogenic mycobacterium. The most extensive genome degradation reported previously was in *Rickettsia prowazekii*, the typhus agent, in which only 76% of the potential coding capacity was used and 12 pseudogenes were identified¹⁶. In comparison with *M. leprae*, the level of gene loss detected was modest, and it is striking that elimination of pseudogenes by deletion lags far behind gene inactivation in both pathogens. Intriguingly, the G+C content of *M. leprae* genes (60.1%) is higher than that of the pseudogenes (56.5%), and the remainder of the genome (54.5%). The high G+C content of *M. leprae*, and other mycobacteria, is apparently driven by the codon preference of active genes, whereas deamination of

cytosines in non-coding regions may have resulted in a more neutral G+C content.

Mosaic organization and horizontal transfer. Although the precise mechanism behind pseudogene formation in *M. leprae* is unclear, the loss of *dnaQ*-mediated proofreading activities of DNA polymerase III (ref. 17) may have contributed. By contrast, there is extensive evidence for large-scale rearrangements and deletions arising from homologous recombination events. Comparison with *M. tuberculosis* delineated ~65 segments that show synteny but differ in their relative order and distribution in the *M. leprae* genome. Breaks in synteny generally correspond to dispersed repeats, transfer RNA genes or gene-poor regions. Copies of all three principal repeats, RLEP, REPPEP and LEPREP, occur at the junctions of discontinuity, suggesting that the mosaic arrangement of the *M. leprae* genome reflects multiple recombination events between related repetitive sequences. In some cases, aberrant recombination may have occurred as truncated repeats exist.

Although there is little sequence similarity indicating that they are insertion sequences, RLEP is probably capable of transposition as it exists within sequences corresponding to known genes. Unlike *M. tuberculosis* H37Rv, which contains at least 2 prophages and 56 intact or truncated insertion sequence (IS) elements^{8,18}, *M. leprae* has only 3 phage-like genes, all with *M. tuberculosis* orthologues, and 26 transposase gene fragments. We did, however, detect one possible case of horizontal transfer of genetic material when we examined the aminoacyl-tRNA synthetase genes. With one exception, all of these are more closely related to *M. tuberculosis* enzymes than to those of any other organism.

Unexpectedly, prolyl-tRNA synthetase, encoded by *proS*, is more similar to the enzymes of *Borrelia burgdorferi* and eukaryotes such as *Drosophila*, humans and yeast. It has been proposed that horizontal transfer of tRNA synthetase genes occurs frequently, and that the pathogen *B. burgdorferi* may have acquired *proS* from its host¹⁹. Comparing the genetic context provides further support for this hypothesis, as the *M. leprae proS* is both displaced and inverted with respect to the *M. tuberculosis* genome (Fig. 2), consistent with a recent acquisition.

Multigene families. Half of the genes (52%) present in *M. tuberculosis* arose from gene-duplication events leading to extensive functional redundancy⁹. Many of these are involved in lipid metabolism or belong to the PE and PPE families, encoding unusual glycine-rich proteins of repetitive structure and unknown function. The latter are confined to certain mycobacterial species²⁰, and represent sources of genetic, and possibly antigenic, variation⁸. The corresponding 167 genes are exceptionally (G+C)-rich and occupy more than 8% of the *M. tuberculosis* genome²¹. By contrast, only 9 intact PE and PPE genes were found in *M. leprae* although 30 pseudogenes were present. No intact members of the PE-PGRS (with multiple Gly-Gly-Ala repeats) subfamily were found. This reduction partly contributes to the smaller genome size and the lower G+C content of *M. leprae*.

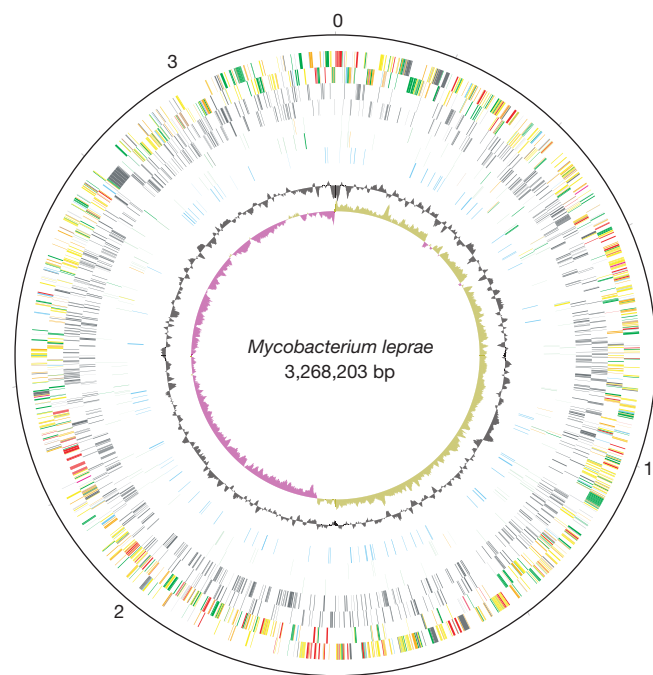


Figure 1 Circular genome map showing the position and orientation of known genes, pseudogenes and repetitive sequences. From the outside: circles 1 and 2 (clockwise and anticlockwise) genes on the – and + strands, respectively; circles 3 and 4, pseudogenes; 5 and 6, *M. leprae* specific genes; 7, repeat sequences; 8, G+C content; 9, G/C bias (G+C)/(G–C). Genes are colour coded using the following functional categories⁸: lipid metabolism (dark grey); intermediary metabolism and respiration (yellow); information pathways (red); regulatory proteins (light blue); conserved hypothetical proteins (orange); proteins of unknown function (light green); insertion sequences and phage related functions (pink); stable RNAs (dark blue); cell wall and cell processes (green); PE and PPE protein families (magenta); virulence, detoxification, adaptation (brown). See http://www.sanger.ac.uk/Projects/M_leprae or <http://genolist.pasteur.fr/Leproma/> for additional information about gene functions. The scale in Mb is indicated (numbers on the outside of the genome).

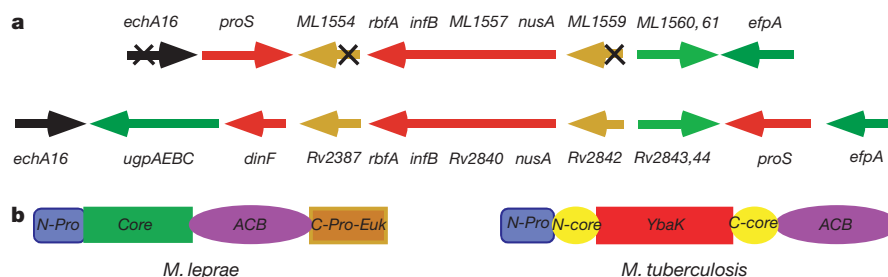


Figure 2 Comparison of the *proS* loci of *M. leprae* and *M. tuberculosis*. **a**, The *M. leprae proS* region is shown above that of *M. tuberculosis*. Genes or operons are depicted by arrows; crosses denote pseudogenes. Note the absence of *ugpAEB* and *dinF* from

M. leprae and the presence of *proS* at this site. **b**, Domain structures of prolyl-tRNA synthetases of bacterial (*M. tuberculosis*) or eukaryotic (*M. leprae*) types after ref. 19. Distinct subdomains are depicted as different shapes.

Table 2 Selected examples of metabolic streamlining

<i>M. tuberculosis</i> Gene	<i>M. leprae</i> Gene	<i>M. leprae</i> Pseudogene	Function	Pathway
<i>gltA1, gltA2, citA</i>	<i>gltA2</i>	<i>citA</i>	Citrate synthase	Krebs cycle
<i>icd1, icd2</i>	<i>icd2</i>	<i>icd1</i>	Isocitrate dehydrogenase	Krebs cycle
<i>icl, aceA</i>	<i>aceA</i>	<i>icl</i>	Isocitrate lyase	Glyoxylate cycle
<i>gnd1, gnd2</i>	<i>gnd1</i>	<i>gnd2</i>	6-phosphogluconate dehydrogenase	Pentose phosphate pathway
<i>pfkA, pfkB</i>	<i>pfkA</i>	<i>pfkB</i>	Phosphofructokinase	Glycolysis
<i>aceE, lpdA, pdhA, pdhB, pdhC, lpd (Rv0462)</i>	<i>aceE, lpd (Rv0462)</i>	<i>lpdA, pdhA, pdhB, pdhC</i>	Pyruvate dehydrogenase complex	Energy metabolism
<i>lldD1, lldD2</i>	<i>lldD2</i>	<i>lldD1</i>	L-lactate dehydrogenase	Respiration
<i>mmaA1, mmaA2, mmaA3, mmaA4, cmaA1, cmaA2, umaA1, umaA2</i>	<i>mmaA1, mmaA4, cmaA2, umaA2</i>	<i>mmaA2, mmaA3, cmaA1, umaA1</i>	Methyltransferase	Mycolic-acid modification
<i>glnA1, glnA2, glnA3, glnA4</i>	<i>glnA1, glnA2</i>	<i>glnA3, glnA4</i>	Glutamine synthase	Glutamine biosynthesis
<i>metA, metB, metC, metE, methH, metK, metZ</i>	<i>metA, metB, metE, methH, metK, metZ</i>	<i>metC</i>	Various	Methionine biosynthesis
<i>bfrA, bfrB</i>	<i>bfrA</i>	<i>bfrB</i>	Bacterioferritin	Iron storage
<i>ligA, ligB, ligC</i>	<i>ligA</i>	<i>ligB</i>	DNA ligase	DNA metabolism

Some PE-PGRS proteins have been shown to be upregulated in *Mycobacterium marinum* during granuloma formation in frogs²². However, this effect is probably not mediated directly by the PE-PGRS, as granulomas are a prominent cytological feature of all forms of leprosy. Essentially all of the gene families⁹ in *M. leprae* have retracted and may now encode 'just enough' activity to permit intracellular growth. Selected examples are given in Table 2 whereas the comprehensive comparison presented in Fig. 3 shows shrinkage of all functional categories.

Metabolic clues

Successive generations of microbiologists have failed to grow *M. leprae* in axenic culture, leading to the notion that the bacterium lacks certain biosynthetic pathways. Complete genome comparisons shed new light on this. Lipid metabolism is prominent in the biochemical repertoire of *M. leprae* but to a lesser extent than in the tubercle bacillus, whose cell envelope has a greater diversity of lipids, glycolipids and carbohydrates²³.

Envelope biogenesis. Mycolic acids are structural components of all mycobacteria and include the alpha mycolates, which lack oxygen functions, and the oxygenated keto- and methoxy-forms.

Reappraisal of mycolic-acid modification is now possible with the reduced *cmaA*, *mmaA* and *umaA* gene sets encoding the effector methyltransferases found in *M. leprae*. *Mycobacterium leprae* contains no methoxy-mycolates²³, probably because it has lost the MmaA2 and MmaA3 enzymes that attach the methoxy group in *M. tuberculosis*^{10,24}. The mycolic acids also have cyclopropane functions²⁵, which in *M. tuberculosis* are carried out by MmaA2 and CmaA1. As both the *mmaA2* and *cmaA1* genes have decayed in *M. leprae*, cyclopropanation must be encoded by one of the related *umaA* genes. Both *umaA2* and *cmaA2* have been shown to be essential for the cyclopropanation function in *M. tuberculosis*²⁶. The same enzymes also catalyse cyclopropanation in *M. leprae*, as their duplicate copies are both inactive (Table 2).

Foremost among the outer lipids of the leprosy bacillus is phenolic glycolipid 1 (PGL1), an envelope component not found in *M. tuberculosis*²⁷. PGL1 is derived from phthiocerol-dimycoserolate (PDIM), an esterified compound lipid generated by myco-cerotic-acid synthase and a type I polyketide synthase (PKS), by addition of three O-methylated deoxy sugars²³. However, we could not detect the genes for the glycosyltransferases, which modify PDIM to produce PGL1, despite extensive comparisons. PDIM, a

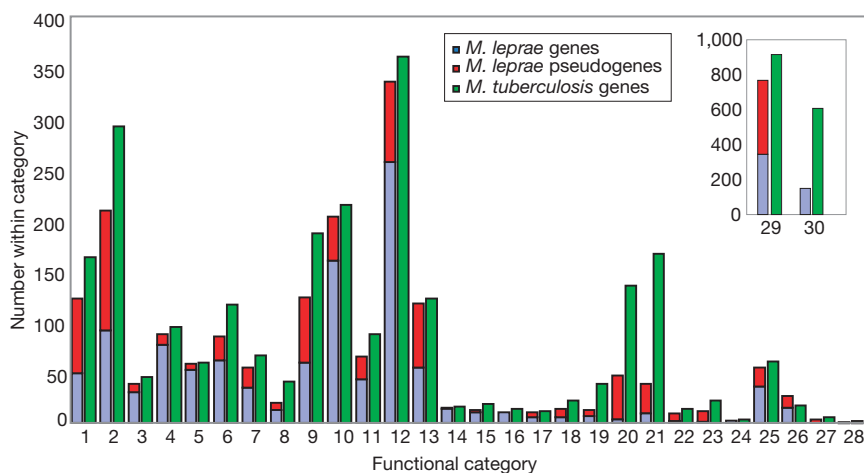


Figure 3 Distribution of genes by functional category. The number of complete (blue) and pseudogenes (red) within each category for *M. leprae* is shown. Data for *M. tuberculosis* (green) were taken from the published genome sequence⁸. Functional categories: 1, small-molecule catabolism; 2, energy metabolism; 3, central intermediary metabolism; 4, amino-acid biosynthesis; 5, nucleoside and nucleotide biosynthesis and metabolism; 6, biosynthesis of cofactors, prosthetic groups and carriers; 7, lipid biosynthesis; 8, polyketide and non-ribosomal peptide synthesis; 9, proteins performing regulatory functions; 10, synthesis and modification of macromolecules; 11, degradation of macromolecules; 12, cell-envelope constituents; 13, transport/binding proteins;

14, chaperones/heat-shock proteins; 15, cell-division proteins; 16, protein and peptide secretion; 17, adaptations and atypical conditions; 18, detoxification; 19, virulence determinants; 20, IS elements and phage-derived proteins; 21, PE and PPE families; 22, antibiotic production and resistance; 23, cytochrome P450 enzymes; 24, coenzyme F420-dependent enzymes; 25, miscellaneous transferases; 26, miscellaneous phosphatases, lyases and hydrolases; 27, cyclases; 28, chelataes. Inset, y axis shows the number of genes within each functional category; x axis shows the functional categories: 29, conserved hypothetical proteins; 30, hypothetical proteins that share no significant similarity with any protein currently in the databases.

virulence factor in *M. tuberculosis*, requires the RND protein, MmpL7, for its transport across the cytoplasmic membrane^{9,28,29}. Of the 18 PKS systems identifiable in *M. tuberculosis*⁸, only 6 were predicted in *M. leprae*; and the number of *mmpL* genes (often linked to PKS genes) is only 5, as opposed to 16 in *M. tuberculosis*, possibly because these genes are no longer required for polyketide or lipid export. Deletion of such systems may be reflected in the lack of mycolipenic and hydroxylipenic acids—polyketides esterified to trehalose in *M. tuberculosis*. Further PKS genes missing from *M. leprae* include the *mbt* operon required for production of the salicylate-based mycobactin siderophores. Lipids, polyketides and aromatic compounds are often substrates for cytochrome-P450 monooxygenases³⁰, enzymes that are exceptionally abundant in *M. tuberculosis*⁸. Astonishingly, none of these is functional in *M. leprae*, although a new member of the P450 family is present.

Lipolysis. Intracellular mycobacteria probably derive much of their energy from the degradation of host-derived lipids³¹, a process initiated by lipases. In contrast to the 22 *lip* genes of *M. tuberculosis*, *M. leprae* has only 2 lipase genes, of which *lipG* clusters with *mmaA* genes and might, therefore, effect fatty-acid remodelling. This appears to leave just one lipase for scavenging fatty acids. In addition to the multifunctional FadA and FadB enzymes, which catalyse β -oxidation, *M. tuberculosis* has numerous alternative systems for fatty-acid degradation⁸. Once again, *M. leprae* has roughly one-third as many potential enzymes; however, there are three-times more FadD acyl-CoA synthases than there are FadE acyl-CoA dehydrogenases, whereas these are predicted in equal amounts in *M. tuberculosis*. This may be explained by the dual role of FadD in β -oxidative and anabolic processes, whereas FadE only participates catabolically.

The acetyl-CoA produced by β -oxidation, or glycolysis, flows into the central pathways of carbon metabolism in *M. leprae*. However, the pattern of 'just enough' genes for each step is firmly established, so that the redundancy seen in *M. tuberculosis* almost never occurs. For instance, there is only one isocitrate lyase (with low predicted activity) capable of participating in the glyoxylate shunt (Table 2)^{32,33}, and one enzyme complex that oxidatively decarboxylates pyruvate to acetyl-CoA, compared with two such systems in *M. tuberculosis*. In the Krebs cycle, as in glycolysis, replicate genes for the same activity are deleted although differences in expression levels might compensate for some missing copies. Thus, although lack of *pdh* genes is reflected in a low rate of oxidative decarboxylation of pyruvate, isocitrate dehydrogenase activity is comparable in host-grown leprosy and tubercle bacilli³⁴ even though a duplicate *icd* gene is inactivated in *M. leprae*.

Central and energy metabolism. Despite an active glyoxylate cycle, there seem to be fundamental differences elsewhere in anaplerotic pathways between *M. leprae* and *M. tuberculosis*. Here, phosphoenolpyruvate (PEP) carboxylase replaces the pyruvate carboxylase of *M. tuberculosis*, and the malic enzyme, which is associated with fast growth in mycobacteria³⁵, is missing. The metabolic implications are that flux between C₃ and C₄ compounds and the balance between glycolysis and gluconeogenesis will be very different. Another missing link between by-products of lipid metabolism and the Krebs cycle is the production of succinyl-CoA by catabolic acetyl/propionyl CoA carboxylases predicted for *M. tuberculosis*⁸.

Other carbon sources lost to *M. leprae* are acetate, because *ackA*, *pta* and *acs* are all inactive, and galactose—the cell-wall galactan can be produced only from glucose because the *galK* and *galT* genes are missing. This might imply that *M. leprae* is limited to growth on a restricted, or even a specialized, combination of carbon sources, on which it can maintain balanced carbon metabolism. Although a similar range of potential substrates is available to both *M. leprae* and *M. tuberculosis* in the host, marked differences in their ability to exploit them are apparent on examination of the systems involved in carbon and nitrogen compound degradation: there are fewer oxidoreductases, oxygenases and short-chain alcohol dehydro-

genases, and their probable regulatory genes (Fig. 3). The inescapable conclusion is that catabolism in *M. leprae* is severely limited.

In the same vein, the leprosy bacillus has lost anaerobic and microaerophilic electron transfer systems, such as formate dehydrogenase, nitrate and fumarate reductase, together with the biosynthetic and transport systems required to produce the cognate prosthetic groups. Likewise, the aerobic respiratory chain of *M. leprae* is truncated, as only the extreme 3' end of the NADH oxidase operon, *nuoA-N*, remains. The consequences of this event are far-reaching, because not only has the potential to produce ATP from the oxidation of NADH been lost, but also the regeneration of NAD⁺ may be limited, as *M. leprae* must rely heavily on *ndh*, which is involved only in recycling NAD⁺. Alternatively, *M. leprae* may oxidize NADH by either diverting pyruvate to acetate and CO₂ using lactate dehydrogenase and lactate oxidase, or diverting PEP to malate or fumarate through oxaloacetate using its PEP carboxylase (an enzyme not found in the tubercle bacillus), which catalyses the reaction only in this direction. Given the loss of genes reviewed above, the acids produced by these two processes cannot be recycled and must be excreted.

Anabolism. In contrast, all the anabolic pathways seem to be relatively intact. With few exceptions, complete enzyme systems are predicted for synthesis of amino acids, purines, pyrimidines, nucleosides, nucleotides, most vitamins and cofactors. This suggests that the availability of these metabolites in phagosomes is either highly limited or that they cannot be transported. It also sets the biology of the leprosy bacillus apart from that of the other obligate parasites for which genomes have been sequenced^{15,16}.

Mycobacterium leprae may, however, be auxotrophic for methionine as *metC*, encoding cystathionine β -lyase, is a pseudogene, whereas the other counterparts of *M. tuberculosis met* genes are all intact. This requirement for methionine may be dictated by the inactivation of the sulphate transport operon, *cysTWA*, implying that *M. leprae* may depend on an organic source of sulphur. Cobinamide auxotrophy is also predicted, as examination of the *cob* genes shows selective deletion of those required for its synthesis, whereas the genes needed to produce vitamin B12 from cobinamide are retained.

Pathogenesis and disease control

The ability to obtain iron is central to a successful pathogenic lifestyle. *Mycobacterium leprae* has many genes for haem and iron-based proteins and uses the iron regulatory systems *ideR* and *furB*. Compared with *M. tuberculosis*, however, *M. leprae* may be severely handicapped as it appears to have lost the *mbt* operon, which encodes the non-ribosomal peptide synthase required for production of the iron-scavenging siderophores mycobactin/exochelin^{8,36,37}. Part of the iron uptake system is functional in *M. leprae*, however, as it transports exochelinMN from *Mycobacterium neoaurum* but not those of *Mycobacterium smegmatis* or *M. tuberculosis*³⁸. The genes for exochelinMN are unknown and seem unlikely to be present in *M. leprae*.

As might be expected given the differences in their respective pathologies, *M. leprae* contains several enzymes that have no counterparts in the tubercle bacillus, including a eukaryotic-like uridine phosphorylase and adenylate cyclase. In addition, there are two transport systems that may have significant physiological roles: an ABC-transporter for sugars; and a second Nrampl-like protein³⁹, could be involved in divalent metal ion uptake. The latter may have been acquired to ensure adequate intracellular iron concentrations resulting from lack of mycobactin siderophores.

M. leprae shows a marked tropism for myelin-producing Schwann cells, and a surface-exposed laminin-binding protein (LBP) of relative molecular mass 21,000 (*M*_r 21K) may be an important virulence factor^{40–42}. Inspection of the genome sequence revealed a single LBP gene, and this also occurs in *M. tuberculosis*. We did not detect any other candidates for virulence genes, and

many of those present in *M. tuberculosis* have been inactivated or lost, including three of the Mce operons encoding putative invasins^{9,43}. Although the leprosy and tubercle bacilli both survive within macrophages, *M. leprae* has no catalase-peroxidase⁴⁴, and fewer peroxidoxins and epoxide hydrolases to combat reactive oxygen species. It has, however, retained both superoxide dismutases.

Elimination of leprosy as a public health problem will not only require continued multidrug therapy but also improved detection of infected individuals so that treatment can be implemented earlier, thus reducing contagion and limiting neurological complications. Diagnosis is difficult in patients with few lesions but, thanks to genomics and the identification of potentially specific proteins, new avenues are now open for developing immunodiagnostic tests. Finally, development of new tuberculosis drugs and vaccines should also benefit directly from our comparative genomic analysis through definition of the core mycobacterial genes. □

Methods

The whole genome sequence was obtained from a combination of sequenced cosmids⁴⁵ and 54,000 end sequences (giving 7.1× coverage) from a pUC18 genomic shotgun library using dye terminator chemistry on ABI373 or ABI377 automated sequencers. The sequences of cosmids previously generated by multiplex sequencing⁴⁶ were used for scaffolding purposes only. The sequence was assembled using Phrap (P. Green, unpublished), finished using GAP4 (ref. 47), and compared with sequences present in public databases using FASTA, BLASTN and BLASTX⁴⁸. Potential coding sequences were predicted, and gene and protein sequences analysed as described previously^{8,49}, using Artemis⁵⁰ to collate data and facilitate annotation. We compared the genome and proteome sequences of *M. leprae* and *M. tuberculosis* H37Rv pairwise to identify conserved genes using the Artemis Comparison Tool (K. Rutherford, unpublished; <http://www.sanger.ac.uk/Software/ACT/>). Pseudogenes had one or more mutations that would ablate expression and were pinpointed by direct comparison with *M. tuberculosis*. A relational database presenting the findings is available (<http://genolist.pasteur.fr/Leproma/>).

Received 10 October; accepted 5 December 2000.

- Karonga Prevention Trial Group. Randomised controlled trial of single BCG, repeated BCG, or combined BCG and killed *Mycobacterium leprae* vaccine for prevention of leprosy and tuberculosis in Malawi. *Lancet* **348**, 17–24 (1996).
- Nordeen, S. K. & Hombach, J. M. in *Tropical Disease Research: Progress 1991–1992. Eleventh Programme Report of the UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases* (eds Walgate, R. & Simpson, K.) 47–55 (World Health Organization, Geneva, 1993).
- World Health Organization in *WHO Weekly Epidemiological Record* **73**, 40 (1998).
- Hansen, G. H. A. Undersøgelser angående spedalskhedens aarsager. *Norsk Magazin for Laegvidenskaben* **4** (Suppl.), 1–88 (1874).
- Kirchheimer, W. K. & Storrs, E. E. Attempts to establish the armadillo (*Dasypus novemcinctus* Linn.) as a model for the study of leprosy. I. Report of lepromatoid leprosy in an experimentally infected armadillo. *Int. J. Lepr.* **39**, 693–702 (1971).
- Franzblau, S. Drug susceptibility testing of *Mycobacterium leprae* in the BACTEC 460 system. *Antimicrob. Agents Chemother.* **33**, 2115–2117 (1989).
- Shepherd, C. C. in *Leprosy* (ed. Hastings, R. C.) 269–286 (Churchill Livingstone, Edinburgh, 1985).
- Cole, S. T. et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544 (1998).
- Tekaia, F. et al. Analysis of the proteome of *Mycobacterium tuberculosis* in silico. *Tubercle Lung Disease* **79**, 329–342 (1999).
- Brosch, R., Gordon, S. V., Eiglmeier, K., Garnier, T. & Cole, S. T. Comparative genomics of the leprosy and tubercle bacilli. *Res. Microbiol.* **151**, 135–142 (2000).
- Philipp, W., Schwartz, D. C., Telenti, A. & Cole, S. T. Mycobacterial genome structure. *Electrophoresis* **19**, 573–576 (1998).
- Stinear, T. P., Jenkin, G. A., Johnson, P. D. R. & Davies, J. K. Comparative genetic analysis of *Mycobacterium ulcerans* and *Mycobacterium marinum* reveals evidence of recent divergence. *J. Bacteriol.* **182**, 6322–6330 (2000).
- Marques, M. A. M., Chitale, S., Brennan, P. J. & Pessolani, M. C. V. Mapping and identification of the major cell-wall associated components of *Mycobacterium leprae*. *Infect. Immun.* **66**, 2625–2631 (1998).
- Jungblut, P. R. et al. Comparative proteome analysis of *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG strains: towards functional genomics of microbial pathogens. *Mol. Microbiol.* **33**, 1103–1117 (1999).
- Andersson, J. O. & Andersson, S. G. E. Insights into the evolutionary process of genome degradation. *Curr. Opin. Genet. Dev.* **9**, 664–671 (1999).
- Andersson, S. G. E. et al. The complete genome sequence of the obligate intracellular parasite *Rickettsia prowazekii*. *Nature* **396**, 133–140 (1998).
- Mizrahi, V., Dawes, S. S. & Rubin, H. in *Molecular Genetics of Mycobacteria* (eds Hatfull, G. F. & Jacobs, W. R. Jr) 159–172 (ASM, Washington DC, 2000).
- Gordon, S. V., Heym, B., Parkhill, J., Barrell, B. & Cole, S. T. New insertion sequences and a novel repeated sequence in the genome of *Mycobacterium tuberculosis* H37Rv. *Microbiology* **145**, 881–892 (1999).

- Wolf, Y. I., Aravind, L., Grishin, N. V. & Koonin, E. V. Evolution of amino-acyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res.* **9**, 689–710 (1999).
- Poulet, S. & Cole, S. T. Repeated DNA sequences in mycobacteria. *Arch. Microbiol.* **163**, 79–86 (1995).
- Cole, S. T. Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv. *FEBS Lett.* **452**, 7–10 (1999).
- Ramakrishnan, L., Federspiel, N. A. & Falkow, S. Granuloma-specific expression of mycobacterium virulence proteins from the glycine-rich PE-PGRS family. *Science* **288**, 1436–1439 (2000).
- Daffe, M. & Draper, P. The envelope layers of mycobacteria with reference to their pathogenicity. *Adv. Microb. Physiol.* **39**, 131–203 (1998).
- Yuan, Y. & Barry, C. E. III A common mechanism for the biosynthesis of methoxy and cyclopropyl mycolic acids in *Mycobacterium tuberculosis*. *Proc. Natl Acad. Sci. USA* **93**, 12828–12833 (1996).
- Draper, P., Dobson, G., Minnikin, D. E. & Minnikin, S. M. The mycolic acids of *Mycobacterium leprae* harvested from experimentally infected nine-banded armadillos. *Ann. Microbiol. (Paris)* **133**, 39–47 (1982).
- Glickman, M. S., Cox, J. S. & Jacobs, W. R. Jr A novel mycolic acid cyclopropane synthetase is required for coding, persistence, and virulence of *Mycobacterium tuberculosis*. *Mol. Cell* **5**, 717–727 (2000).
- Melancon-Kaplan, J. et al. Immunological significance of the cell wall of *Mycobacterium leprae*. *Proc. Natl Acad. Sci. USA* **85**, 1917–1921 (1988).
- Cox, J. S., Chen, B., McNeil, M. & Jacobs, W. R. Jr Complex lipid determines tissue-specific replication of *Mycobacterium tuberculosis* in mice. *Nature* **402**, 79–83 (1999).
- Camacho, L. R., Ensergueix, D., Perez, E., Gicquel, B. & Guilhot, C. Identification of a virulence gene cluster of *Mycobacterium tuberculosis* by signature-tagged transposon mutagenesis. *Mol. Microbiol.* **34**, 257–267 (1999).
- Peterson, J. A. & Graham, S. E. A close family resemblance: the importance of structure in understanding cytochromes P450. *Structure* **6**, 1079–1085 (1998).
- Wheeler, P. R. & Ratledge, C. in *Tuberculosis: Pathogenesis, Protection, and Control* (ed. Bloom, B. R.) 353–385 (Am. Soc. Microbiol., Washington DC, 1994).
- Honer, Z. B., Miczak, A., Swenson, D. L. & Russell, D. G. Characterization of activity and expression of isocitrate lyase in *Mycobacterium avium* and *Mycobacterium tuberculosis*. *J. Bacteriol.* **181**, 7161–7167 (1999).
- McKinney, J. D. et al. Persistence of *Mycobacterium tuberculosis* in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase. *Nature* **406**, 735–738 (2000).
- Wheeler, P. R. Oxidation of carbon sources through the tricarboxylic acid cycle in *Mycobacterium leprae* grown in armadillo liver. *J. Gen. Microbiol.* **130**, 381–389 (1984).
- Ratledge, C. R. in *The Biology of the Mycobacteria* (eds Ratledge, C. & Stanford, J.) 53–94 (Academic, San Diego, 1982).
- De Voss, J. J. et al. The salicylate-derived mycobactin siderophores of *Mycobacterium tuberculosis* are essential for growth in macrophages. *Proc. Natl Acad. Sci. USA* **97**, 1252–1257 (2000).
- Quadri, L. E., Sello, J., Keating, T. A., Weinreb, P. H. & Walsh, C. T. Identification of a *Mycobacterium tuberculosis* gene cluster encoding the biosynthetic enzymes for assembly of the virulence-conferring siderophore mycobactin. *Chem. Biol.* **5**, 631–645 (1998).
- Hall, R. M. & Wheeler, P. R. Exochelin-mediated iron uptake into *Mycobacterium leprae*. *Int. J. Lepr. Other Mycobact. Dis.* **51**, 490–494 (1983).
- Makui, H. et al. Identification of the *Escherichia coli* K-12 Nmp orthologue (MntH) as a selective divalent metal ion transporter. *Mol. Microbiol.* **35**, 1065–1078 (2000).
- Shimoi, Y., Ng, V., Matsumura, K., Fischetti, V. A. & Rambukkana, A. A 21-kDa surface protein of *Mycobacterium leprae* binds peripheral nerve laminin-2 and mediates Schwann cell invasion. *Proc. Natl Acad. Sci. USA* **96**, 9857–9862 (1999).
- Rambukkana, A. Role of alpha-dystroglycan as a Schwann cell receptor for *Mycobacterium leprae*. *Science* **282**, 2076–2079 (1998).
- Rambukkana, A., Szelzer, J. L., Yurchenco, P. D. & Tuomanen, E. I. Neural targeting of *Mycobacterium leprae* mediated by the G domain of the laminin-α2 chain. *Cell* **88**, 811–821 (1997).
- Arruda, S., Bomfim, G., Knights, R., Huima-Byron, T. & Riley, L. W. Cloning of an *M. tuberculosis* DNA fragment associated with entry and survival inside cells. *Science* **261**, 1454–1457 (1993).
- Eiglmeier, K., Fsihi, H., Heym, B. & Cole, S. T. On the catalase-peroxidase gene, *katG*, of *Mycobacterium leprae* and the implications for treatment of leprosy with isoniazid. *FEMS Microbiol. Lett.* **149**, 273–278 (1997).
- Eiglmeier, K., Honoré, N., Woods, S. A., Caudron, B. & Cole, S. T. Use of an ordered cosmid library to deduce the genomic organisation of *Mycobacterium leprae*. *Mol. Microbiol.* **7**, 197–206 (1993).
- Smith, D. R. et al. Multiplex sequencing of 1.5 Mb of the *Mycobacterium leprae* genome. *Genome Res.* **7**, 802–819 (1997).
- Bonfield, J. K., Smith, K. F. & Staden, R. A new DNA sequence assembly program. *Nucleic Acids Res.* **24**, 4992–4999 (1995).
- Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. Issues in searching molecular sequence databases. *Nature Genet.* **6**, 119–129 (1994).
- Parkhill, J. et al. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* **404**, 502–505 (2000).
- Rutherford, K. Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944–945 (2000).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or from http://www.sanger.ac.uk/Projects/M_leprae, or as paper copy from the London editorial office of Nature.

Acknowledgements

We are grateful to B. R. Bloom, P. J. Brennan, M. J. Colston, J. Grosset and B. Ji for advice, reagents and encouragement. This work was supported by the New York Community Trust, ILEP, the Association Française Raoul Follereau, the Wellcome Trust and the Institut Pasteur.

Correspondence and requests for materials should be addressed to S.T.C. (e-mail: stcole@pasteur.fr). The EMBL databank accession number for the sequence is AL450380.

A dusty torus around the luminous young star LkH α 101

Peter G. Tuthill*, John D. Monnier† & William C. Danchi‡

* Astronomy Department, School of Physics, University of Sydney, Sydney NSW 2006, Australia

† Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA

‡ NASA Goddard Space Flight Center, Infrared Astrophysics, Code 685, Greenbelt, Maryland 20771, USA

A star forms when a cloud of dust and gas collapses. It is generally believed that this collapse first produces a flattened rotating disk^{1,2}, through which matter is fed onto the embryonic star at the centre of the disk. When the temperature and density at the centre of the star pass a critical threshold, thermonuclear fusion begins. The remaining disk, which can still contain up to 0.3 times the mass of the star^{3–5}, is then sculpted and eventually dissipated by the radiation and wind from the newborn star. But this picture of the structure and evolution of the disk remains speculative because of the lack of morphological data of sufficient resolution and uncertainties regarding the underlying physical processes. Here we present images of a young star, LkH α 101, in which the structure of the inner accretion disk is resolved. We find that the disk is almost face-on, with a central gap (or cavity) and a hot inner edge. The cavity is bigger than previous theoretical predictions⁶, and we infer that the position of the inner edge is probably determined by sublimation of dust grains by direct stellar radiation, rather than by disk-reprocessing or viscous-heating processes as usually assumed²⁹.

The Herbig Ae/Be stars, thought to be pre-main-sequence stars of intermediate mass, have generated considerable recent controversy over their circumstellar structure. Accretion-disk models were found to fit the spectral energy distribution (SED) provided that an optically thin hole or cavity was allowed around the central star^{6,7}. This model drew almost immediate criticism over missing accretion luminosity⁸ and problems with forbidden emission line profiles not matching expectations⁹. Further confusion arose from the finding that the SED could be fitted by spherically symmetric circumstellar shells^{10,11} or composite shell–disk models¹².

LkH α 101 is amongst the brightest young stellar objects in the near-infrared, despite its location behind considerable line-of-sight obscuration (recent estimates of visible extinction A_V range from 9.4 (ref. 13) to 18.5 (ref. 14)). Interferometric observations with the Keck I telescope capable of imaging structure on scales of tens of milli-arcseconds¹⁵ have enabled us to observe the circumstellar environment of LkH α 101 in the near-infrared at an unprecedented level of detail.

Images presented in Fig. 1, taken in the H and K bands, show the bright core to be resolved into a circular limb-brightened disk with a marked asymmetric brightening along the southwest side. A region of suppressed flux, or dark hole, is apparent at the centre. LkH α 101 is also revealed as a binary star, with a companion at a separation of 180 mas to the E–NE in the 1.65 μ m map. This binary component is still visible in the map at 2.27 μ m, although it is now much fainter, appearing only at the lowest contour with a signal-to-noise ratio around 1. The relative brightening of the companion from K to H band implies a significantly hotter apparent spectrum.

The circular appearance of the resolved primary component implies a face-on viewing angle onto an accretion disk. Fitting to the images in Fig. 1 gives a maximum possible inclination angle of about 35° to the line-of-sight. The central star is not seen at either

wavelength, as its contribution is swamped by the extremely high infrared luminosity of the disk. By examining the visibility curves, we find an upper limit of 5% for any unresolved component in the maps. This should not be surprising as the infrared excesses exhibited by many Herbig Ae/Be stars imply that circumstellar material completely dominates the SED from the infrared through to the radio wavelengths.

The depression or hole in the emission that we saw at the centre of the ring provides definitive confirmation of the scenario of an accretion disk with a central optically thin cavity in this young Herbig Ae/Be system. Although this finding is in accord with the scenario of ref. 6, this agreement may be superficial as the cavity observed here is much larger than those considered previously.

The asymmetric flux resulting in the bright crescent appearance to the southwest is a result of the disk being tilted away from the line-of-sight in this direction. This allows us to see through the central cavity to the hot inner edge of the disk on the far side (in the southwest), while in the north-east our view is partially blocked by cooler regions of the disk in the line-of-sight. Just such a “peculiar horseshoe-like feature” was well predicted⁷ from models with simulated disks tilted at an angle of 30°, albeit in the mid-infrared ($\lambda = 12 \mu$ m) wavelengths.

In order to make meaningful comparisons with literature models, we need to estimate the basic properties of the central star driving LkH α 101. The extremely high absolute bolometric luminosity of $4.8 \times 10^4 L_\odot$ (ref. 13), where $L_\odot$ is the luminosity of the Sun, some 75% of which is thought to come from a deeply-embedded cluster containing hundreds of objects¹⁶, was based on an earlier distance determination of 800 pc (ref. 17). However, a recent distance estimate mainly based on radio-photometry points to 160 pc (ref. 18)—typical of the Taurus–Auriga star formation complex with which LkH α 101 may be associated. If this distance is correct, the implied intrinsic luminosity (a factor of ~ 25 times lower) is $4.8 \times 10^2 L_\odot$ —more appropriate for a mid–late B star than an O star, as was supposed. However, radio observations showing an ionized shell^{19,20} imply Lyman continuum photon fluxes such as would result from a much hotter early B star (B0 to B1). Given the uncertainties in the extinction and the known global anisotropies, an exact classification of LkH α 101 is problematic, but assuming it to be an early B-type zero age main sequence (ZAMS) star gives an approximate mass $M_\star$ of 5–10 $M_\odot$, an effective temperature $T_{\text{eff}} \approx 20,000$ K and a radius $R_\star \sim 4R_\odot$ (ref. 21). (Here $M_\odot$ and $R_\odot$ are respectively the mass and the radius of the Sun.)

The circular ridge of brightest emission has a radius of 21 mas measured from the K-band map of Fig. 1. If we identify this as the hot edge of an inwardly terminated accretion disk⁷, this implies a physical radius of 3.4 AU for a distance of 160 pc (or alternatively 16.8 AU at 800 pc). This is vastly larger than the $3 < R_{\text{hole}}/R_\star < 25$ inner cavities required by ref. 6 to fit the near-infrared inflections in the SED. For example, models of early B stars from this sample had derived cavity radii of 0.3–0.6 AU, around an order of magnitude smaller than the one we see. Although we find some qualitative agreement, we conclude that our images are inconsistent with classical accretion disk models.

This reinforces findings from recently reported high-resolution observations of the Herbig objects with the Infrared Optical Telescope Array (IOTA)^{22,23} and the Palomar Testbed Interferometer²⁴. Although most of the visibility data^{22,23} could be fitted with standard viscously heated accretion disks ($T \propto R^{-3/4}$), an unlikely set of extreme inclination angles was required. Instead, simpler spherical gaussian or ring models worked better, suggesting that the ensemble of observations were most consistent with spherical shell geometries. The one-sided banana-shaped brightening on our images does not suggest spherical symmetry. The IOTA, with limited sky rotation (typically $\sim 30^\circ$ baseline rotation) and Fourier coverage may

not have detected similar asymmetries if present in the targets of refs 22, 23. The match between the asymmetric limb-brightened structure we see and the models of ref. 7 supports models based on the idea of an accretion disk with a central cavity for LkH α 101. Resolved imaging of more systems will be required to settle the conflict between spherical versus disk models for Herbig Ae/Be stars.

There is a rapidly developing theoretical framework for interpreting young stellar objects at later epochs where the winds and radiation from the fully luminous central star sculpt the remnant accretion disk by photo-evaporation and wind ram pressure. By equating the plasma sound speed with the keplerian orbital velocity, Hollenbach *et al.*⁵ defined the gravitational radius r_g beyond which material is no longer gravitationally bound and will escape as an ionized wind. Following ref. 5, we find for a $7.5M_\odot$ star that $r_g = 50$ AU, although this does entail extrapolation to less massive stars than were originally considered (alternatively, $r_g = 100$ AU if the heavier $15M_\odot$ stellar mass¹⁶ is adopted). This is significantly larger than the apparent structure we see, implying that the resolved disk fits well within the picture of a gravitationally bound circumstellar structure.

Photo-evaporative models⁷ were constructed for an $8.4M_\odot$ star surrounded by a $1.6M_\odot$ circumstellar disk—not dissimilar to our expectations for LkH α 101. Radiative and hydrodynamical modelling resulted in inner disk cavities of around 40 AU in radius. Thus the model scale is larger than we observed by almost an order of magnitude, although this discrepancy would vanish if the 800-pc distance scale is adopted. However, the visual match of the appearance of disk models inclined at 30° (ref. 7) and the observed limb-

brightened crescent of Fig. 1 encourage hope that the basic structure and physics of radiant and wind processes describing the sculpting and illumination of the stellar disk are on the right track. Perhaps self-consistent models can be found with closer inner radii.

We proceed with a far simpler comparison of our 3.4-AU cavity with the radius at which dust sublimates as a result of heating from the stellar radiation field: $R_s = \frac{1}{2} \sqrt{Q_R(T_\star/T_s)^2 R_\star}$ where T_s is the sublimation temperature and $Q_R = Q(T_\star)/Q(T_s)$ is the ratio of the absorption efficiencies $Q(T)$ of the dust for radiation at colour temperature T of the incident and re-emitted field. Assuming that grains surviving at the inner edge will be those best able to radiate away the energy from the star (blackbody grains) we assumed $Q_R = 1$. Taking $T_s = 1,500$ K and using the stellar parameters above gives $R_s = 1.7$ AU—a reasonable match to our observed cavity given the uncertainties in the stellar properties. Also, our fit was to the peak of the ridge of emission in Fig. 1, but the innermost edge may be somewhat smaller than this. Finding the dust at (or beyond) the sublimation radius is analogous to similar findings for asymptotic giant branch stars, implying that the same basic radiative equilibrium process may set the inner cavity size in both cases.

We turn now to the binary companion which was found to contribute some 0.8% of the flux at K and 4.0% of the flux at H. Such a luminous (relatively) blue object may be the source of Lyman-continuum photoionizing flux. Alternatively the (unseen) object at the centre of the disk may power the energetic emission-line spectrum. Given the flux ratios in the near-infrared, it is possible to estimate the temperature of the companion relative to the disk. Unfortunately, the uncertainties in the extinction mentioned above hamper the de-reddening calculations, so we can only

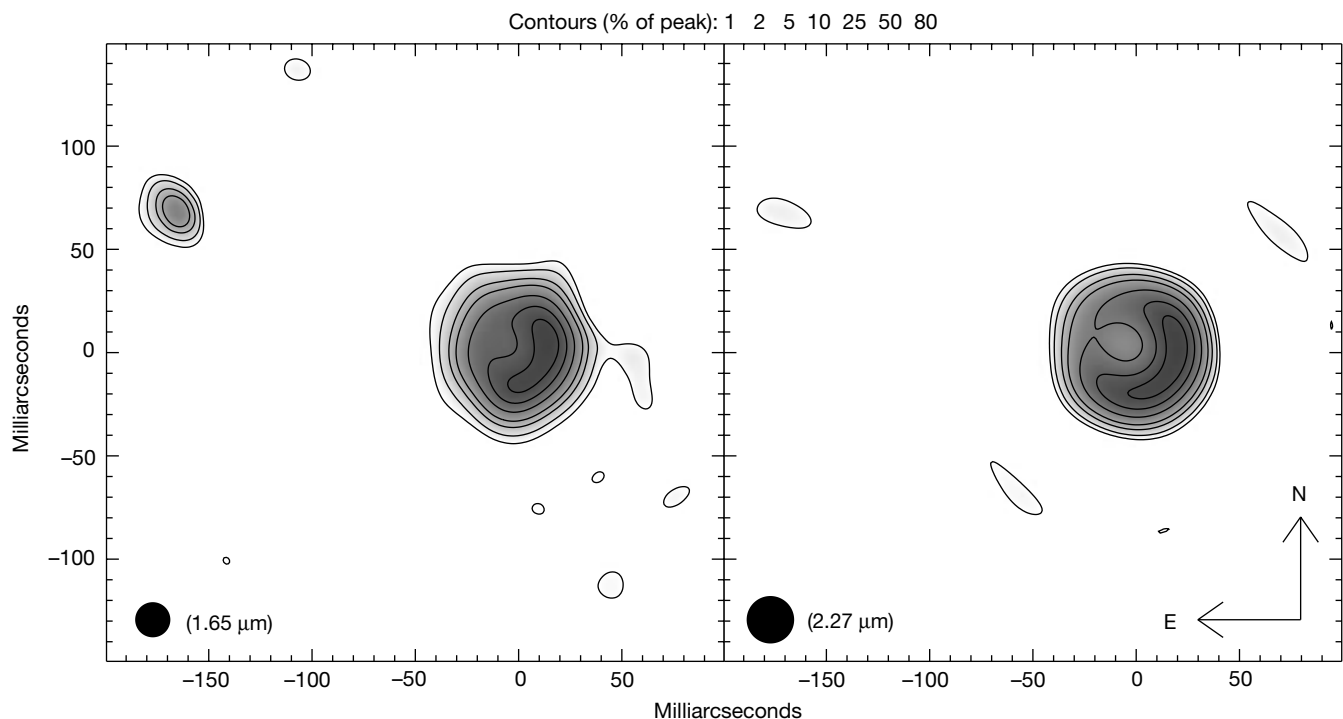


Figure 1 Maps of the LkH α 101 system. Maximum-entropy image reconstructions at 1.65 (left) and 2.27 μm (right) from data taken in September 1998 (JD2451086). Peak contour has darkest shading. Filter bandwidths were 0.33 and 0.16 μm respectively. As it is difficult to extract both absolute astrometry and photometry with our interferometric techniques, the origin has been centred on the disk and images scaled to the brightest pixel. The K-band map is of slightly superior quality to the H-band owing to the less-detrimental impact of the seeing at longer wavelengths. High-spatial-resolution information was extracted from high-magnification (20 mas per pixel) fast-readout (~ 130 ms) data frames taken with an annulus-shaped aperture mask¹⁵. Consecutive sets

of 100 images, interleaved between LkH α 101 and nearby point-source objects, allowed calibration of the statistical behaviour of the telescope-atmosphere point-spread function. Data were processed to extract Fourier bispectral data, with the final output being a calibrated set of 630 Fourier amplitudes and 7,140 closure phases. High-quality images were produced from an algorithm based on the maximum entropy method^{25,26}, although other methods such as CLEAN have also been used with near-identical results. The fidelity of our mapping process has been extensively tested on binary stars and other known compact asymmetric targets^{27,28}.

place limits, finding the companion to be at least 2,000 K hotter than the disk.

We detected the secondary at multiple epochs and the components show no apparent relative motion greater than 15 mas yr^{-1} . As it seems reasonable to assume that the two objects do form a bound pair, assuming a total system mass of $\sim 10 M_{\odot}$ results in an orbital period of 49 yr, given a physical separation of 28.8 AU. The apparent motion exhibited by such a system is comparable to our limit, and so extra observations at future epochs should reveal the orbit of this system. We hope such future observations, and the extension of imaging to shorter wavelengths, will help to identify which component provides the ionizing flux powering this young active region. $\square$

Received 3 October; accepted 28 December 2000.

1. Terebey, S., Shu, F. H. & Cassen, P. The collapse of the cores of slowly rotating isothermal clouds. *Astrophys. J.* **286**, 529–551 (1984).
2. Yorke, H. W., Bodenheimer, P. & Laughlin, G. The formation of protostellar disks. 2: Disks around intermediate-mass stars. *Astrophys. J.* **443**, 199–208 (1995).
3. Cassen, P. M., Smith, B. F., Miller, R. H. & Reynolds, R. T. Numerical experiments on the stability of preplanetary disks. *Icarus* **48**, 377–392 (1981).
4. Shu, F. H., Tremaine, S., Adams, F. C. & Ruden, S. P. Sling amplification and eccentric gravitational instabilities in gaseous disks. *Astrophys. J.* **358**, 495–514 (1990).
5. Hollenbach, D., Johnstone, D., Lizano, S. & Shu, F. Photoevaporation of disks around massive stars and application to ultracompact H II regions. *Astrophys. J.* **428**, 654–669 (1994).
6. Hillenbrand, L. A., Strom, S. E., Vrba, F. J. & Keene, J. Herbig Ae/Be stars—Intermediate-mass stars surrounded by massive circumstellar accretion disks. *Astrophys. J.* **397**, 613–643 (1992).
7. Kessel, O., Yorke, H. W. & Richling, S. Photoevaporation of protostellar disks. III. The appearance of photoevaporating disks around young intermediate mass stars. *Astron. Astrophys.* **337**, 832–846 (1998).
8. Hartmann, L., Kenyon, S. J. & Calvet, N. The excess infrared emission of Herbig Ae/Be stars—Disks or envelopes? *Astrophys. J.* **407**, 219–231 (1993).
9. Böhm, T. & Catala, C. Forbidden lines in Herbig Ae/Be stars: the [O I](1F) 6300.31 Å and 6363.79 Å lines. I. Observations and qualitative analysis. *Astron. Astrophys.* **290**, 167–175 (1994).
10. Miroshnichenko, A., Ivezic, Z. & Elitzur, M. On protostellar disks in Herbig Ae/Be stars. *Astrophys. J.* **475**, L41–L44 (1997).
11. Pezzuto, S., Strafella, F. & Lorenzetti, D. On the circumstellar matter distribution around Herbig Ae/Be stars. *Astrophys. J.* **485**, 290–307 (1997).
12. Miroshnichenko, A., Ivezic, Z., Vinkovic, D. & Elitzur, M. Dust emission from Herbig Ae/Be stars: evidence for disks and envelopes. *Astrophys. J.* **529**, L115–L118 (1999).
13. Barsony, M., Scoville, N. Z., Schombert, J. M. & Claussen, M. J. The circumstellar environment of the emission-line star LkH-alpha 101. *Astrophys. J.* **362**, 674–690 (1990).
14. Hou, J., Jiang, D. & Fu, C. Physical properties of stellar winds from young stellar objects. *Astron. Astrophys.* **327**, 725–735 (1997).
15. Tuthill, P. G., Monnier, J. D., Danchi, W. C., Wishnow, E. H. & Haniff, C. A. Michelson interferometry with the Keck I telescope. *Publ. Astron. Soc. Pacif.* **112**, 555–565 (2000).
16. Barsony, M., Schombert, J. M. & Kis-Halas, K. The LkH-alpha 101 infrared cluster. *Astrophys. J.* **379**, 221–231 (1991).
17. Herbig, G. H. The spectrum of LkH-alpha 101 in the near-infrared. *Astrophys. J.* **169**, 537–541 (1971).
18. Stine, P. C. & O'Neal, D. Radio emission from young stellar objects near LkH-alpha 101. *Astron. J.* **116**, 890–894 (1998).
19. Harris, S. 5 GHz radio observations of LkH-alpha 101, M1-82#1 and other infrared sources. *Mon. Not. R. Astron. Soc.* **174**, 601–607 (1976).
20. Cohen, M., Bieging, J. H. & Schwartz, P. R. VLA observations of mass loss from T Tauri stars. *Astrophys. J.* **253**, 707–715 (1982).
21. Panagia, N. Some physical parameters of early-type stars. *Astron. J.* **78**, 929–934 (1973).
22. Millan-Gabet, R. et al. Sub-astronomical unit structure of the near-infrared emission from AB Aurigae. *Astrophys. J.* **513**, L131–L134 (1999).
23. Millan-Gabet, R., Schloerb, F. P. & Traub, W. A. Spatially resolved circumstellar structure of Herbig Ae/Be stars in the near-infrared. *Astrophys. J.* **546**, 358–381 (2001).
24. Akeson, R. L. et al. Infrared interferometric observations of young stellar objects. *Astrophys. J.* **543**, 313–317 (2000).
25. Gull, S. F. & Skilling, J. Maximum entropy method in image processing. *IEEE Proc. F* **131**, 646–650 (1984).
26. Sivia, D. S. *Phase Extension Methods*. Thesis, Cambridge Univ. (1987).
27. Tuthill, P. G., Monnier, J. D. & Danchi, W. C. A dusty pinwheel nebula around the massive star WR 104. *Nature* **398**, 487 (1999).
28. Tuthill, P. G., Monnier, J. D., Danchi, W. C. & Lopez, B. Smoke signals from IRC + 10216: 1. Milliarcsecond proper motions of the dust. *Astrophys. J.* **543**, 284 (2000).
29. Lynden-Bell, D. & Pringle, J. E. The evolution of viscous discs and the origin of the nebular variables. *Mon. Not. R. Astron. Soc.* **168**, 603–637 (1974).

Acknowledgements

We would like to thank D. Sivia for the maximum-entropy mapping program VLBMEM. Data were obtained at the W.M. Keck Observatory, made possible by the generous support of the W.M. Keck Foundation, operated as a scientific partnership among the California Institute of Technology, the University of California and NASA. This work was supported through grants from the National Science Foundation.

Correspondence and requests for materials should be addressed to P.T. (e-mail: gekko@physics.usyd.edu.au).

Experimental entanglement distillation and 'hidden' non-locality

Paul G. Kwiat^{*†}, Salvador Barraza-Lopez^{*†}, André Stefanov[‡] & Nicolas Gisin[‡]

^{*} Physics Division, P-23, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

[‡] Group of Applied Physics, University of Geneva, 1211 Geneva 4, Switzerland

Entangled states are central to quantum information processing, including quantum teleportation¹, efficient quantum computation² and quantum cryptography³. In general, these applications work best with pure, maximally entangled quantum states. However, owing to dissipation and decoherence, practically available states are likely to be non-maximally entangled, partially mixed (that is, not pure), or both. To counter this problem, various schemes of entanglement distillation, state purification and concentration have been proposed^{4–11}. Here we demonstrate experimentally the distillation of maximally entangled states from non-maximally entangled inputs. Using partial polarizers, we perform a filtering process to maximize the entanglement of pure polarization-entangled photon pairs generated by spontaneous parametric down-conversion^{12,13}. We have also applied our methods to initial states that are partially mixed. After filtering, the distilled states demonstrate certain non-local correlations, as evidenced by their violation of a form of Bell's inequality^{14,15}. Because the initial states do not have this property, they can be said to possess 'hidden' non-locality^{6,16}.

The basic idea of distillation is to extract from an ensemble of pairs of non-maximally entangled qubits a smaller number of pairs with a higher degree of entanglement. In the literature and the scientific community this procedure has also been described as 'purification' and 'concentration'. We adopt the term 'distillation', following refs 7 and 9 and others, as most descriptive of the nature of the process. Following ref. 11, we reserve 'purification' for processes which increase purity, but not entanglement; and 'concentration' for processes which increase both.

Because the qubits may be separated in space, only local operations and classical communication (LOCC) are allowed. The simplest procedure—Bernstein's 'procrustean method'⁴—is as follows. A two-qubit state of the form

$$|\phi_{\epsilon}\rangle = (\epsilon|00\rangle + |11\rangle)/\sqrt{1+\epsilon^2}, \quad (1)$$

which is non-maximally entangled for $\epsilon \neq 1$, can be transformed into the maximally entangled state $|\phi^{+}\rangle = |00\rangle + |11\rangle$ (unnormalized), simply by subjecting one of the qubits to a generalized measurement^{17–19} that takes $|0\rangle \rightarrow |0\rangle$ and $|1\rangle \rightarrow \epsilon|1\rangle$. This nonunitary filtering process equalizes the contribution of the two terms in equation (1), thereby yielding perfect entanglement. Such a procedure is the necessary first step in a distillation protocol known to apply to all entangled two-qubit states⁷.

Our set-up to investigate procrustean distillation is shown in Fig. 1. As described elsewhere¹³, using two adjacent nonlinear crystals (β -barium borate, BBO), we can readily produce photon pairs in any state of the form $|\phi_{\epsilon}\rangle = (\epsilon|HH\rangle + |VV\rangle)/\sqrt{1+\epsilon^2}$, where H and V represent horizontal and vertical polarization (coding for qubits $|0\rangle$ and $|1\rangle$), respectively. The basic principle of the source is that an incident ultraviolet pump photon may be split into two infrared photons either in the first crystal (from the pump's vertical polarization component), in which case the daughter

[†] Present addresses: University of Illinois, Urbana-Champaign, Department of Physics, 1110 West Green Street, Urbana, Illinois 61801-3080, USA (P.G.K.); Escuela Superior de Física y Matemáticas, IPN, Mexico D.F., 07738, Mexico (S.B.-L.).

photons are both horizontally polarized; or in the second crystal (from the pump's horizontal polarization component), in which case they are vertically polarized. Owing to the coherent nature of the process, and the fact that the output spatial modes of the photons are nearly identical, the photons produced in this way can exhibit almost perfect polarization-entanglement¹². By simply varying the pump polarization, we can arbitrarily set ϵ .

The photons are detected in coincidence using silicon avalanche photodiodes, thereby projecting out the large vacuum state from the possibility of the pump photon not downconverting, and selecting the two-photon contribution to the total quantum mechanical state. (In our experiment the residual contribution from higher-order processes, that is, with simultaneous multiple pairs, is less than 0.1%.) We thus restrict our discussion to this two-photon subspace. The polarization of each photon may be analysed in an arbitrary basis, by means of a quarter wave plate (QWP), half wave plate (HWP), and polarizing beam splitter (PBS) in each arm. By making 16 measurements of the polarization correlations in various bases (for example, HH, HV, 45° V, and so on), which are essentially the two-photon analogues of the classical Stoke's parameters²⁰, we may derive the density matrix describing the polarization state of the pairs produced¹³. Figure 2a (left) shows typical data, corresponding to the state $0.41|HH\rangle + 0.91|VV\rangle$, for which $\epsilon = 0.45$.

The procrustean filtering can be realized experimentally by inserting into one path a series of coated glass slabs, tilted about the vertical axis by 58° (approximately Brewster's angle for these slabs). Owing to the well known polarization-dependent reflectivity²⁰, the transmitted photons are preferentially horizontally polarized. Specifically, with four slabs, the transmission probability for horizontal polarization was $T_H = 0.89$, while for vertical polarization it was only $T_V = 0.18$. The state ρ_{out} after filtering is shown in Fig. 2a (right). Its fidelity with the maximally entangled state $|\phi^+\rangle = (|HH\rangle + |VV\rangle)/\sqrt{2}$ (defined as $\text{Tr}[\langle\phi^+|\rho_{\text{out}}|\phi^+\rangle]$) is 0.99 ± 0.04 , indicating a successful distillation. We performed similar distillation for several values of ϵ , changing our partial polarizers accordingly; in every case the fidelity of the resulting density matrix with the desired maximally-entangled state $|\phi^+\rangle$ was close to 1.

The efficiency of the procrustean method—the probability that a given input pair survives the filtering process—is directly linked to

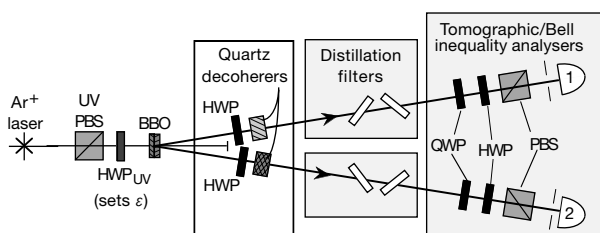


Figure 1 Experimental set-up used to investigate entanglement distillation and hidden non-locality. Spontaneous down-conversion in two adjacent β -barium borate (BBO) crystals results in polarization-entangled photons. The initial degree of entanglement ϵ is determined by the linear polarization angle θ of the pump beam (at 351 nm, from an argon ion laser): $\epsilon = \tan\theta$. In the experiment requiring partial mixture as well, each photon is passed through an adjustable half wave plate (HWP) and a 1-cm-thick quartz element, to introduce decoherence in a particular basis. The final quarter wave plate (QWP) and half wave plate in each arm, along with polarizing beam splitters (PBS), enable analysis of the polarization correlations in any basis, allowing tomographic reconstruction of the density matrix, and measurement of Bell's inequalities. The distillation filters comprise glass substrates, tilted near Brewster's angle, and symmetrically arranged in pairs to compensate for transverse displacements. The result is an adjustable ratio of the transmissions for vertical and horizontal polarization. In the experimental to concentrate entanglement from a non-maximally entangled state, filtering was performed in one arm only. In the experiment on hidden non-locality, identical filters were placed in each arm, as shown.

the value of ϵ , that is, to the required amount of distillation. In an ideal system we would have $T_H = 1$, $T_V = \epsilon^2$, so that the yield of output maximally entangled pairs is $2\epsilon^2/(1 + \epsilon^2)$. In our experiment, the values of T_H were slightly less than one, because of some absorption in the InSnO coating on our glass slabs. Our measured efficiencies are plotted in Fig. 2b, along with the ideal theoretical prediction, and the expectation based on our actual measured values of T_H and T_V . Remarkably, this simplest method of distillation always has a higher yield/input pair than more complicated (and presently physically unrealizable) schemes relying on jointly processing n pairs, as long as $n \leq 5$ (ref. 4): for the particular non-maximally entangled states we studied ($\epsilon = 0.45, 0.68, 0.77$ and 0.90) our observed yields are superior unless $n \geq 5, 10, 9$ and 28 , respectively.

Before coming to the more general case, let us emphasize the difference between the distillation filtering process and the post-selection done by the detectors. The former necessarily depends on the photon polarization to select out a particular subensemble of pairs of photons from the initial total ensemble of pairs; this filtering affects the state, that is, it changes the relative weightings within the two-photon density matrix. In contrast, the detectors merely reveal the two-photon component of the state, by selecting pairs out of an initial superposition of photons and no-photons (that is, the vacuum); this post-selection is made independently of all analysis settings, as the detector efficiencies do not depend on polarization. If this were not so, we could observe apparently non-local effects from completely classical ensembles^{6,16}.

Next consider the situation of an initial state that is partially mixed, in addition to being non-maximally entangled:

$$\rho_{\epsilon,\lambda} = \lambda|\phi^+\rangle\langle\phi^+| + \frac{(1-\lambda)}{2}(|HV\rangle\langle HV| + |VH\rangle\langle VH|) \quad (2)$$

The first term corresponds to a pure, non-maximally entangled component, the second to a mixed component. One feature of

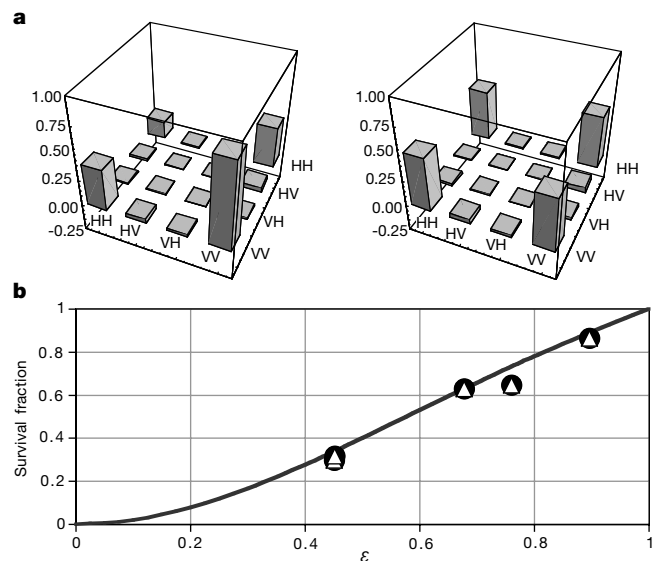


Figure 2 Experimental results showing entanglement distillation of non-maximally entangled states. **a**, Left is the measured two-photon polarization density matrix for the initial, unfiltered state: $0.41|HH\rangle + 0.91|VV\rangle$; right is the resulting maximally-entangled state after an appropriate generalized filtering measurement has been performed. (Only the real parts are shown; the imaginary components, which theoretically are strictly zero, were typically less than 1%.) **b**, The fractional 'survival' probability is plotted as a function of the initial degree of entanglement ϵ . The solid curve is the theoretical prediction under the optimal condition where T_H of the filter is 1. Our data points (solid circles) are well predicted by a theory (open triangles) using our actual measured values of T_H (typically ~ 0.92) and T_V .

Table 1 Summary of distillation data for testing Bell's inequalities

Trial	ϵ_{non}	λ_{non}	$S_{\text{non}}^{\text{calc}}$	$S_{\text{non}}^{\text{exp}}$	$\epsilon_{\text{filtered}}$	$\lambda_{\text{filtered}}$	$S_{\text{filtered}}^{\text{calc}}$	$S_{\text{filtered}}^{\text{exp}}$
1	0.46	0.78	1.63 ± 0.04	1.60 ± 0.02	1.02	0.73	2.02 ± 0.04	2.02 ± 0.03
2	0.47	0.83	1.81 ± 0.04	1.82 ± 0.02	1.01	0.79	2.20 ± 0.04	2.22 ± 0.03
3	0.47	0.87	1.96 ± 0.04	1.94 ± 0.02	0.99	0.83	2.31 ± 0.04	2.34 ± 0.02

The table shows the measured ϵ and λ parameters of the non-filtered and filtered states, and the calculated and experimentally measured S values. The errors in the latter are calculated from Poisson statistics. The calculated predictions for S were obtained by averaging S over an ensemble of density matrices slightly deviated from the ideal state corresponding to ϵ and λ .

entangled quantum systems is that some of the correlations they predict cannot be explained by an local realistic model. This statement is made more quantitative by Bell's inequalities (BI)¹⁴. Here we shall always restrict our discussion to the Clauser–Horne–Shimony–Holt (CHSH) version¹⁵ (and its equivalents), though in general there may exist more revealing tests of non-locality, for example, ones which rely on sequences of measurements on each particle¹⁶. The CHSH inequality places constraints on the value of S , a combination of four polarization correlation probabilities—two possible analysis settings for each photon. If $|S| \leq 2$, no quantum mechanical entanglement is necessary to explain the correlations, that is, some local model can reproduce them. For a maximally-entangled state the maximum value of $S = 2\sqrt{2}$. The optimal analysis settings for a particular state ρ may be determined, for example by using the analytical method of ref. 21. Depending on the values of ϵ and λ , $\rho_{\epsilon,\lambda}$ may be unable to violate any CHSH inequality⁶, for instance, if the state is too mixed ($\lambda \leq 1/\sqrt{2}$). Figure 3 indicates the ranges over which no CHSH BI can be violated.

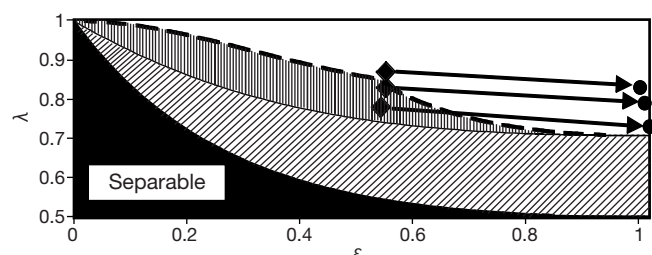


Figure 3 Contour plots in the ϵ – λ plane, showing the curves for $S = 2$, for states of the form (2). The dashed curve corresponds to the unfiltered state; below this line one cannot violate a Bell's inequality of the CHSH-type. The solid curve shows the $S = 2$ threshold for states after an optimal distillation of the filtering sort described here—no CHSH violation is ever possible in the lower, diagonally-hatched region below this curve. The vertically-hatched overlap region indicates states possessing 'hidden' non-locality—a local filtering operation enables one to violate a suitable Bell's inequality. Finally, the solid region at the bottom shows states which are completely separable, that is, no entanglement can ever be recovered by only local operations and classical communication. The diamonds indicate the initial values for our three experimental trials (see Table 1); the circles are the post-distillation values.

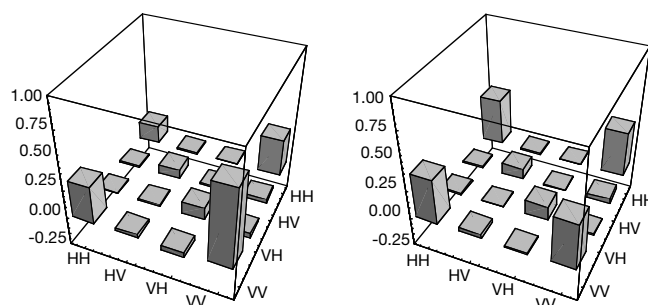


Figure 4 Density matrices for states displaying 'hidden' non-locality. The initial state (left) is $0.83\rho_{0.43\text{HH}+0.90\text{V}} + 0.09(\rho_{\text{HV}} + \rho_{\text{VH}})$. After symmetric filtering, the state is transformed (right) into one which is $0.79\rho_{0.71\text{HH}+0.71\text{V}} + 0.11(\rho_{\text{HV}} + \rho_{\text{VH}})$. We note that the purity,

The process of distillation can alter this, however. The filtering method is similar to that described above, except that now we make a generalized measurement in each arm. Specifically, if each partial polarizer has $T_{\text{H}} = 1$ and $T_{\text{V}} = \epsilon$, the state (2) becomes (up to a normalization factor):

$$\frac{2\epsilon^2\lambda}{1+\epsilon^2}|\phi^+\rangle\langle\phi^+| + \frac{\epsilon(1-\lambda)}{2}(|\text{HV}\rangle\langle\text{HV}| + |\text{VH}\rangle\langle\text{VH}|) \quad (3)$$

We note that the relative contribution of the entangled term has been reduced by a factor of $2\epsilon/(1+\epsilon^2)$ —the filtered state (3) is more mixed than the original $\rho_{\epsilon,\lambda}$. Nevertheless, because the remaining pure component is now maximally entangled, the state may violate a suitable BI (Fig. 3).

The method to experimentally produce generalized states as in state (2) is discussed elsewhere^{22,23}. By sending each photon through a thick birefringent element (~ 1 cm quartz), we introduce a frequency-dependent phase offset between two polarization components. As this offset is greater than the coherence length of the photons (~ 100 μm , in our case determined by the 5-nm (full width at half maximum) interference filter in front of the detector), the resulting polarization quantum state of the pair can be mixed. The birefringent element acts to entangle the polarization and frequency degrees of freedom; when we trace over the frequency (since the detectors are insensitive to wavelength over the collection bandwidth), the reduced density matrix for polarization becomes mixed²³. By appropriately adjusting the pump polarization (with HWP_{UV}) and the decoherence basis (with half wave plates before the quartz), we can prepare states of the form $\rho_{\epsilon,\lambda}$.

We investigated distillation on several different starting states (see Table 1). From a tomographically measured density matrix of each initial, unfiltered state, we analytically and numerically determine QWP and HWP analysis settings predicting the maximum S value. We then measure S directly using these (16) settings, and the normalization procedure introduced in ref. 24. The first trial corresponds to a state which clearly does not violate any CHSH inequality, but is just at the limit when filtered. In the second trial there is a definite transition between no initial violation and violation after filtering (Figs 4a and b show the density matrices before and after filtering, respectively). The final trial starts with a state which is near the threshold of violation, and ends with a strong violation (17σ). To our knowledge, this is the first time Bell's inequalities have been tested using mixed states.

defined as $\text{Tr}[\rho^2]$, is actually reduced from 0.69 to 0.64 by the distillation process, that is, the final state is more mixed than the initial.

Whereas our distillation procedure employs only local operations (filtering) and classical communication (coincidence counting), our non-locality demonstrations clearly rely on ‘conditional probabilities’^{16,25}—we include only events in which both photons are transmitted by their respective partial polarizers. Nevertheless, we emphasize that these filters are before the analysers, and are thus independent of the polarization measurements. The filtering process is used to select a particular subensemble of the original states. The entanglement for members of this subensemble (corresponding to coincident detections) is higher than the average for the entire ensemble (which includes photons reflected at the partial polarizers—these possess no entanglement whatsoever), and therefore they are able to demonstrate non-local effects, such as violations of Bell’s inequalities. This is completely consistent with the fact that LOCC cannot be used to increase the average entanglement of an ensemble⁴. For instance, for our second trial, the initial degree of entanglement (of formation)²⁶ before filtering is $E = 0.32$; after filtering the transmitted photons have $E = 0.44$, but the reflected photons have $E = 0$. □

Received 11 September; accepted 15 December 2000.

- Bennet, C. H. *et al.* Teleporting an unknown quantum state via dual classical and Einstein–Podolsky–Rosen channels. *Phys. Rev. Lett.* **70**, 1895–1899 (1992).
- Bennett, C. & DiVincenzo, D. P. Quantum information and computation. *Nature* **404**, 247–255 (2000).
- Ekert, A. K. Quantum cryptography based on Bell theorem. *Phys. Rev. Lett.* **67**, 661–663 (1991).
- Bennett, C. H., Bernstein, H. J., Popescu, S. & Schumacher, B. Concentrating partial entanglement by local operations. *Phys. Rev. A* **53**, 2046–2052 (1996).
- Bennett, C. H. *et al.* Purification of noisy entanglement and faithful teleportation via noisy channels. *Phys. Rev. Lett.* **76**, 722–725 (1996).
- Gisin, N. Hidden quantum nonlocality revealed by local filters. *Phys. Lett. A* **210**, 151–156 (1996).
- Horodecki, M., Horodecki, P. & Horodecki, R. Inseparable two spin-1/2 density matrices can be distilled to a singlet form. *Phys. Rev. Lett.* **78**, 574–577 (1997).
- Linden, N., Massar, S. & Popescu, S. Purifying noisy entanglement requires collective measurements. *Phys. Rev. Lett.* **81**, 3279–3282 (1998).
- Opatrny, T. & Kurizki, G. Optimization approach to entanglement distillation. *Phys. Rev. A* **60**, 167–172 (1999).
- Bose, S., Vedral, V. & Knight, P. L. Purification via entanglement swapping and conserved entanglement. *Phys. Rev. A* **60**, 194–197 (1999).
- Thew, R. T. & Munro, W. J. Entanglement manipulation and concentration. *Phys. Rev. A* (in the press); also preprint quant-ph/0012049 at (xxx.lanl.gov) (2000).
- Kwiat, P. G., Waks, E., White, A. G., Appelbaum, L. & Eberhard, P. H. Ultrabright source of polarization-entangled photons. *Phys. Rev. A* **60**, R773–R776 (1999).
- White, A. G., James, D. F. V., Eberhard, P. H. & Kwiat, P. G. Nonmaximally entangled states: Production, characterization, and utilization. *Phys. Rev. Lett.* **83**, 3103–3107 (1999).
- Bell, J. S. On the Einstein–Podolsky–Rosen paradox. *Physics* **1**, 195 (1965).
- Clauser, J., Horne, M., Shimony, S. & Holt, R. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.* **23**, 880–884 (1969).
- Popescu, S. Bell’s inequalities and density matrices: Revealing ‘hidden’ nonlocality. *Phys. Rev. Lett.* **74**, 2619–2622 (1995).
- Helstrom, C. W. *Quantum Detection and Estimation Theory* (Academic, New York, 1976).
- Huttner, B. *et al.* Unambiguous quantum measurement of non-orthogonal states. *Phys. Rev. A* **54**, 3783–3789 (1996).
- Clark, R. B. M., Chefles, A., Barnett, S. M. & Riis, E. Experimental demonstration of optimal unambiguous state discrimination. *Phys. Rev. A* (in the press); preprint quant-ph/0007063 at (xxx.lanl.gov) (2000).
- Born, M. & Wolf, E. *Principles of Optics* 30–32 (Cambridge Univ. Press, Cambridge, 1999).
- Horodecki, R., Horodecki, P. & Horodecki, M. Violating Bell inequality by mixed spin-1/2 states: Necessary and sufficient condition. *Phys. Lett. A* **200**, 340–344 (1995).
- Kwiat, P. G., Berglund, A. J., Altepeter, J. B. & White, A. G. Experimental verification of decoherence-free subspaces. *Science* **290**, 498–501 (2000).
- Berglund, A. J. *Quantum Coherence and Control in One- and Two-Photon Optical Systems*. Thesis, Dartmouth College B.A. (2000); also preprint quant-ph/0010001 at (xxx.lanl.gov) (2000).
- Aspect, A., Grangier, P. & Roger, G. Experimental realization of Einstein–Podolsky–Rosen–Bohm gedankenexperiment: A new violation of Bell’s inequalities. *Phys. Rev. Lett.* **49**, 91–94 (1982).
- Zukowski, M., Horodecki, R., Horodecki, M. & Horodecki, P. Generalized quantum measurements and local realism. *Phys. Rev. A* **58**, 1694–1698 (1998).
- Wootters, W. K. Entanglement of formation of an arbitrary state of two qubits. *Phys. Rev. Lett.* **80**, 2245–2248 (1998).

Acknowledgements

We thank H. Bernstein, D. DiVincenzo, B.-G. Englert, L. Hardy, D. James and M. Zukowski for helpful discussions. This work was supported in part by the National Security Agency (NSA) and Advanced Research and Development Activity (ARDA), and by the European IST FET QuComm project.

Correspondence and requests for materials should be addressed to P.G.K. (e-mail: Kwiat@uiuc.edu).

Fluid particle accelerations in fully developed turbulence

A. La Porta, Greg A. Voth, Alice M. Crawford, Jim Alexander & Eberhard Bodenschatz

Laboratory of Atomic and Solid State Physics, Laboratory of Nuclear Studies, Cornell University, Ithaca, New York 14853-2501, USA

The motion of fluid particles as they are pushed along erratic trajectories by fluctuating pressure gradients is fundamental to transport and mixing in turbulence. It is essential in cloud formation and atmospheric transport^{1,2}, processes in stirred chemical reactors and combustion systems³, and in the industrial production of nanoparticles⁴. The concept of particle trajectories has been used successfully to describe mixing and transport in turbulence^{3,5}, but issues of fundamental importance remain unresolved. One such issue is the Heisenberg–Yaglom prediction of fluid particle accelerations^{6,7}, based on the 1941 scaling theory of Kolmogorov^{8,9}. Here we report acceleration measurements using a detector adapted from high-energy physics to track particles in a laboratory water flow at Reynolds numbers up to 63,000. We find

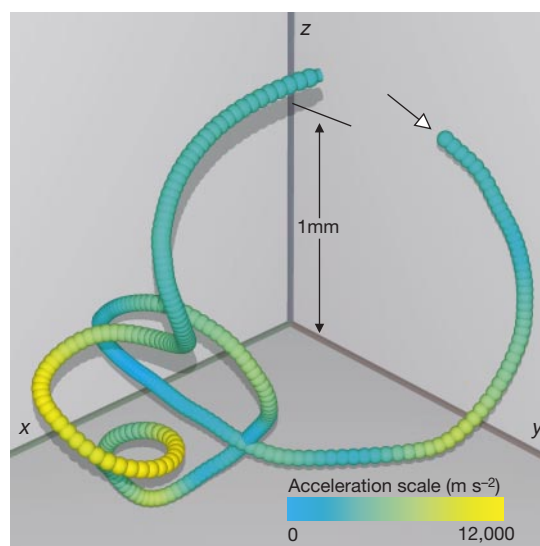


Figure 1 Measured particle trajectory. The three-dimensional time-resolved trajectory of a 46- μm -diameter particle in a turbulent water flow at Reynolds number 63,000 ($R_\lambda = 970$). A sphere marks the measured position of the particle in each of 300 frames taken every 0.014 ms ($\approx \tau_p/20$). The shading indicates the acceleration magnitude, with the maximum value of 12,000 m s^{-2} corresponding to approximately 30 standard deviations. The turbulence is generated between coaxial counter-rotating disks^{24,25} in a closed-flow chamber of volume 0.1 m^3 with rotation rates ranging from 0.15 Hz to 7.0 Hz, giving r.m.s. velocity fluctuation $\bar{u}$ in the range $0.018 \text{ m s}^{-1} < \bar{u} < 0.87 \text{ m s}^{-1}$. Measurements are made in an 8- mm^3 volume at the centre of the apparatus where the mean velocity is zero and the flow is nearly homogeneous but not isotropic. As a result of a mean stretching of the flow along the propeller axis the r.m.s. fluctuations are one-third larger for the transverse velocity components than for the axial component. The energy dissipation was determined from measurements of the transverse second-order structure function D_{NN} and the Kolmogorov relation $D_{NN} = \frac{4}{3} C_1 (\epsilon r)^{2/3}$ with $C_1 = 2.13$ (ref. 26), ϵ the turbulent energy dissipation, and r the particle separation. The dissipation was found to be related to the r.m.s. velocity fluctuation by $\epsilon = \bar{u}^3/L$ with an energy injection scale $L = (71 \pm 7) \text{ mm}$. Using the definition of the Taylor microscale Reynolds number $R_\lambda = (15\bar{u}L/\nu)^{1/2}$, the range of Reynolds numbers accessible is $140 \leq R_\lambda \leq 970$ (in terms of the classical Reynolds number $1,300 \leq \text{Re} \leq 63,000$). At the highest Reynolds number the system is characterized by Kolmogorov distance and timescales of $\eta = 18 \mu\text{m}$ and $\tau_\eta = 0.3 \text{ ms}$, respectively.

that, within experimental errors, Kolmogorov scaling of the acceleration variance is attained at high Reynolds numbers. Our data indicate that the acceleration is an extremely intermittent variable—particles are observed with accelerations of up to 1,500 times the acceleration of gravity (equivalent to 40 times the root mean square acceleration). We find that the acceleration data reflect the anisotropy of the large-scale flow at all Reynolds numbers studied.

In principle, fluid particle trajectories are easily measured by seeding a turbulent flow with minute tracer particles and following their motions with an imaging system. In practice this can be a very challenging task because we must fully resolve particle motions which take place on timescales of the order of the Kolmogorov time, $\tau_\eta = (\nu/\epsilon)^{1/2}$ where ν is the kinematic viscosity, ϵ is the turbulent energy dissipation and η is the Kolmogorov distance. This is exemplified in Fig. 1, which shows a measured three-dimensional, time-resolved trajectory of a tracer particle undergoing violent accelerations in our turbulent water flow, for which $\tau_\eta = 0.3$ ms. The particle enters the detection volume on the upper right, is pushed to the left by a burst of acceleration and comes nearly to a stop before being rapidly accelerated (at 1,200 times the acceleration of gravity) upward in a corkscrew motion. This trajectory illustrates the difficulty in following tracer particles—a particle's acceleration can go from zero to 30 times its root mean square (r.m.s.) value and back to zero in fractions of a millisecond and within distances of hundreds of micrometres.

Conventional detector technologies are effective for low Reynolds number flows^{10,11}, but do not provide adequate temporal resolution at high Reynolds numbers. However, the requirements are met by the use of silicon strip detectors as optical imaging elements in a particle-tracking system. The strip detectors used in our experiment (see Fig. 2a) were developed to measure particle tracks in the vertex detector of the CLEO III experiment operating at the Cornell Electron Positron Collider¹². When applied to particle tracking in turbulence (see Fig. 2b) each detector measures a one-dimensional projection of the image of the tracer particles. Using a data acquisition system designed for the turbulence experiment, several detectors can be simultaneously read out at up to 70,000 frames per second.

The acceleration of a fluid particle, $\mathbf{a}^+$, in a turbulent flow is given by the Navier–Stokes equation:

$$\mathbf{a}^+ = -\frac{\nabla p}{\rho} + \nu \nabla^2 \mathbf{u} \quad (1)$$

where p is the pressure, ρ is the fluid density, and $\mathbf{u}$ is the velocity

field. In fully developed turbulence the viscous damping term is small compared to the pressure gradient term^{13,14} and therefore the acceleration is closely related to the pressure gradient.

Our measurement of the distribution of accelerations is shown in Fig. 3, where the probability density function of a normalized acceleration component is plotted at three Reynolds numbers. All of the distributions have a stretched exponential shape, in which the tails extend much further than they would for a gaussian distribution with the same variance. This indicates that accelerations many times the r.m.s. value are not as rare as one might expect, that is, the acceleration is an extremely intermittent variable. The acceleration flatness, shown in the inset to Fig. 3, characterizes the intermittency of the acceleration, and would be 3 for a gaussian distribution. These flatness values are consistent with direct numerical simulation (DNS) at low Reynolds number¹⁴ and exceed 60 at the highest Reynolds numbers.

The prediction by Heisenberg and Yaglom of the variance of an acceleration component based on Kolmogorov '41 theory^{8,9} is

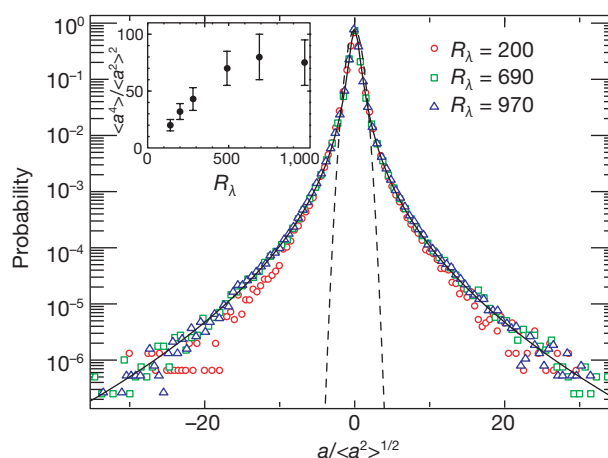


Figure 3 Acceleration distribution. Probability density functions of the transverse acceleration normalized by its standard deviation at different Reynolds numbers. The acceleration is measured from parabolic fits over $0.75\tau_\eta$ segments of each trajectory. The solid line is a parameterization of the highest Reynolds number data using the function $P(a) = C \exp(-a^2/(1 + |a\beta/\sigma|^\gamma)\sigma^2)$, with $\beta = 0.539$, $\gamma = 1.588$ and $\sigma = 0.508$, and the dashed line is a gaussian distribution with the same variance. The inset shows the flatness of the acceleration distribution, $\langle a^4 \rangle / \langle a^2 \rangle^2$, evaluated using $0.5\tau_\eta$ parabolic fits) as a function of R_λ .

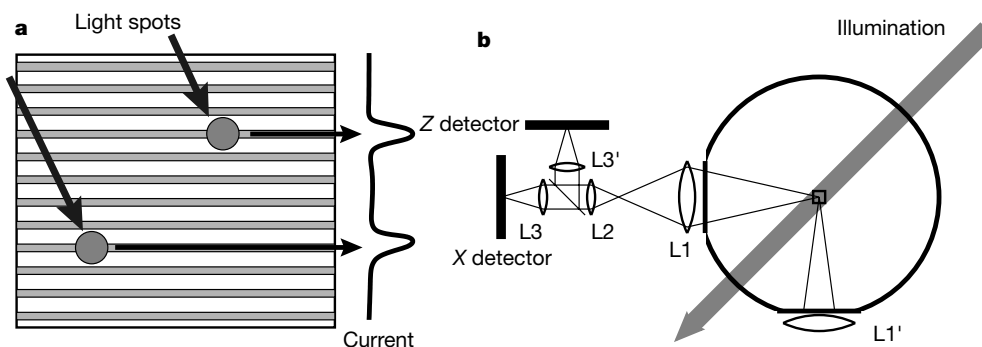


Figure 2 Apparatus. **a**, Schematic representation of the CLEO III strip detector¹², in which grey bars indicate sense strips which collect charge carriers freed by optical radiation. The 511 strips allow measurement of the one-dimensional projection of the light striking the detector. The detector may be read out 70,000 times per second. **b**, A combination of lenses (L1, L2, L3, L3') is used to image the active volume onto a pair of strip detectors which are oriented to measure the x and y coordinates. Another detector assembly may be

placed on the opposite port (L1') to measure y and z. The flow is illuminated by a 6 W argon ion laser beam oriented at 45° with respect to the two viewports. The optics image (46 ± 7) - μ m-diameter transparent polystyrene spheres which have a density of 1.06 g cm^{-3} . Particle positions are measured with an accuracy of 0.1 strips, corresponding to $0.7 \mu\text{m}$ in the flow.

$$\langle a_i a_j \rangle = a_0 \epsilon^{3/2} \nu^{-1/2} \delta_{ij} \quad (2)$$

where a_0 is a universal constant which is approximately 1 in a model assuming gaussian fluctuations^{6,7,13,15}. However, DNS has found that a_0 depends on ϵ . Conventionally this is expressed in terms of the Taylor microscale Reynolds number, R_λ , which is related to the conventional Reynolds number by $R_\lambda = (15\text{Re})^{1/2}$ and is proportional to $\epsilon^{1/6}$. Using this notation, DNS results indicate $a_0 \approx R_\lambda^{1/2}$ for $R_\lambda < 250$ (ref. 14), with a tendency to level off¹⁶ as R_λ approaches 470.

Our measurement of the Kolmogorov constant a_0 is shown in Fig. 4 for eight orders of magnitude of scaling in acceleration variance. We find a_0 to be anisotropic and to depend significantly on the Reynolds number. The a_0 values for both components increase as a function of Reynolds number up to $R_\lambda \approx 500$, above which they are approximately constant. The trend in a_0 is consistent with DNS results in the range $140 \leq R_\lambda \leq 470$ (refs 14, 11, 17). However, the constant value of a_0 at high Reynolds number suggests that Kolmogorov '41 scaling^{8,9} is attained at higher Reynolds numbers. Owing to experimental uncertainties, weak deviations from the Kolmogorov '41 scaling^{8,9} such as the $a_0 \approx R_\lambda^{0.135}$ prediction of the Borgas multifractal model¹⁸ cannot be ruled out by our measurements.

The acceleration variance is larger for the transverse component than for the axial component at all values of the Reynolds number. This is shown in the inset to Fig. 4 where the ratio of the Kolmogorov constants for the axial and transverse acceleration components is plotted as a function of Reynolds number. The anisotropy is large at low Reynolds number and diminishes to a small value at $R_\lambda = 970$. This observation is consistent with recent experimental results which indicate that anisotropy persists to high Reynolds numbers^{19,20}.

In summary, our measurements indicate that the Heisenberg–Yaglom scaling of acceleration variance is observed for $500 \leq R_\lambda \leq 970$. At lower Reynolds number, our measurements are consistent with the nonuniversal scaling observed in DNS^{14,16}. Our measurements show that the anisotropy of the large scales affects the acceleration components even at $R_\lambda \approx 1,000$. It is impossible to say on the basis of these measurements if the

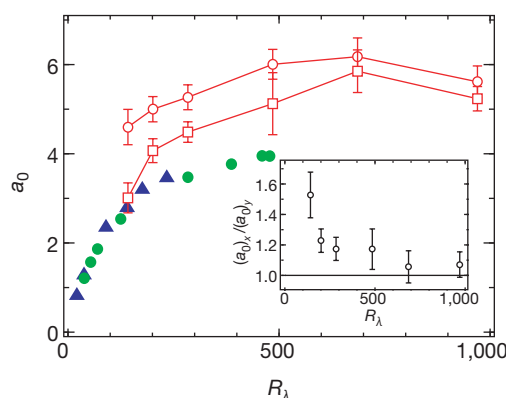


Figure 4 Acceleration a_0 as a function of R_λ . Open red circles indicate a transverse component and open red squares the axial component of the acceleration variance. Direct numerical simulation data is represented by blue triangles¹⁴ and green circles¹⁶. The error bars represent random and systematic errors in the measurement of the acceleration variance. There is an additional uncertainty of 15% in the overall scaling of the vertical axis for the experimental data due to the uncertainty in the measured value of the energy dissipation. The degree to which the 45- μm -diameter tracer particles follow the flow was investigated by measuring the acceleration variance as a function of particle size and density. The results, to be published elsewhere, confirm that the acceleration variance of the 46- μm particles is within a few per cent of the zero particle size limit. The inset shows the ratio of the a_0 values for transverse and axial components of the acceleration.

anisotropy will persist as the Reynolds number approaches infinity. We have found the acceleration to be a very intermittent variable, with extremely large accelerations often arising in structures such as the one shown in Fig. 1.

Our results have immediate application for the development of lagrangian stochastic models, some of which use a_0 directly as a model constant. These models are being developed and used to efficiently simulate mixing, particulate transport, and combustion in practical flows with varying Reynolds numbers^{3,21,22}. Our research also has surprising implications for everyday phenomena. For instance, a mosquito flying on a windy day (wind speed 18 km h⁻¹ and an altitude of 1 m) would experience an r.m.s. acceleration of 15 m s⁻². But given the extremely intermittent nature of the acceleration, the mosquito could expect to experience accelerations of 150 m s⁻² (15 times the acceleration of gravity) every 15 seconds. This may explain why, under windy conditions, a mosquito would prefer to cling to a blade of grass rather than take part in the rollercoaster ride through the Earth's turbulent boundary layer²³. □

Received 6 November 2000; accepted 3 January 2001.

- Vaillancourt, P. A. & Yau, M. K. Review of particle-turbulence interactions and consequences for cloud physics. *Bull. Am. Meteorol. Soc.* **81**, 285–298 (2000).
- Weil, J. C., Sykes, R. I. & Venkatram, A. Evaluating air-quality models: Review and outlook. *J. Appl. Meteorol.* **31**, 1121–1145 (1992).
- Pope, S. B. Lagrangian PDF methods for turbulent flows. *Annu. Rev. Fluid Mech.* **26**, 23–63 (1994).
- Pratsinis, S. E. & Srinivas, V. Particle formation in gases, a review. *Powder Technol.* **88**, 267–273 (1996).
- Shraiman, B. I. & Siggia, E. D. Scalar turbulence. *Nature* **405**, 639–646 (2000).
- Heisenberg, W. Zur statistischen theorie der turbulenz. *Z. Phys.* **124**, 628–657 (1948).
- Yaglom, A. M. On the acceleration field in a turbulent flow. *C.R. Acad. URSS* **67**, 795–798 (1949).
- Kolmogorov, A. N. The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers. *Dokl. Akad. Nauk SSSR* **30**, 301–305 (1941).
- Kolmogorov, A. N. Dissipation of energy in the locally isotropic turbulence. *Dok. Akad. Nauk SSSR* **31**, 538–540 (1941).
- Virant, M. & Dracos, T. 3D PTV and its application on Lagrangian motion. *Meas. Sci. Technol.* **8**, 1539–1552 (1997).
- Ott, S. & Mann, J. An experimental investigation of relative diffusion of particle pairs in three-dimensional turbulent flow. *J. Fluid Mech.* **422**, 207–223 (2000).
- Skubic, P. et al. The CLEO III silicon tracker. *Nucl. Instrum. Meth. A* **418**, 40–51 (1998).
- Batchelor, G. K. Pressure fluctuations in isotropic turbulence. *Proc. Camb. Phil. Soc.* **47**, 359–374 (1951).
- Vedula, P. & Yeung, P. K. Similarly scaling of acceleration and pressure statistics in numerical simulations of isotropic turbulence. *Phys. Fluids* **11**, 1208–1220 (1999).
- Obukhov, A. M. & Yaglom, A. M. The microstructure of turbulent flow. *Prikl. Mat. Mekh.* **15**, (1951); translation TM 1350 (National Advisory Committee for Aeronautics (NACA), Washington DC, 1953).
- Gotoh, T. & Fukayama, D. Pressure spectrum in homogeneous turbulence. *Phys. Rev. Lett.* (submitted).
- Gotoh, T. & Rogallo, R. S. Intermittency and scaling of pressure at small scales in forced isotropic turbulence. *J. Fluid Mech.* **396**, 257–285 (1999).
- Borgas, M. S. The multifractal Lagrangian nature of turbulence. *Phil. Trans. R. Soc. Lond. A* **342**, 379–411 (1993).
- Kurien, S. & Sreenivasan, K. R. Anisotropic scaling contributions to high-order structure functions in high-Reynolds-number turbulence. *Phys. Rev. E* **62**, 2206–2212 (2000).
- Shen, X. & Warhaft, Z. The anisotropy of the small scale structure in high Reynolds number ($R_\lambda \sim 1000$) turbulent shear flow. *Phys. Fluids* **12**, 2976–2989 (2000).
- Reynolds, A. M. A second-order Lagrangian stochastic model for particle trajectories in inhomogeneous turbulence. *Q. J. R. Meteorol. Soc.* **125**, 1735–1746 (1999).
- Sawford, B. L. & Yeung, P. K. Eulerian acceleration statistics as a discriminator between Lagrangian stochastic models in uniform shear flow. *Phys. Fluids* **12**, 2033–2045 (2000).
- Bidlingmeyer, W. L., Day, J. F. & Evans, D. G. Effect of wind velocity on suction trap catches of some Florida mosquitos. *J. Am. Mosquito Contr.* **11**, 295–301 (1995).
- Voth, G. A., Satyanarayan, K. & Bodenschatz, E. Lagrangian acceleration measurements at large Reynolds numbers. *Phys. Fluids* **10**, 2268–2280 (1998).
- La Porta, A., Voth, G. A., Moisy, F. & Bodenschatz, E. Using cavitation to measure statistics of low-pressure events in large-Reynolds-number turbulence. *Phys. Fluids* **12**, 1485–1496 (2000).
- Sreenivasan, K. R. On the universality of the Kolmogorov constant. *Phys. Fluids* **7**, 2778–2784 (1995).

Acknowledgements

This research is supported by the Physics Division of the National Science Foundation. We thank R. Hill, M. Nelkin, S. B. Pope, E. Siggia, and Z. Warhaft for stimulating discussions and suggestions throughout the project. We also thank C. Ward, who assisted in the initial development of the strip detector. E.B. and A.L.P. are grateful for support from the Institute of Theoretical Physics at the University of California, Santa Barbara.

Correspondence and requests for materials should be addressed to E.B. (e-mail: eb22@cornell.edu).

Prediction of absolute crystal-nucleation rate in hard-sphere colloids

Stefan Auer & Daan Frenkel

FOM Institute for Atomic and Molecular Physics, Kruislaan 407, 1098 SJ Amsterdam, The Netherlands

Crystal nucleation is a much-studied phenomenon, yet the rate at which it occurs remains difficult to predict. Small crystal nuclei form spontaneously in supersaturated solutions, but unless their size exceeds a critical value—the so-called critical nucleus—they will re-dissolve rather than grow. It is this rate-limiting step that has proved difficult to probe experimentally. The crystal nucleation rate depends on P_{crit} , the (very small) probability that a critical nucleus forms spontaneously, and on a kinetic factor (κ) that measures the rate at which critical nuclei subsequently grow. Given the absence of *a priori* knowledge of either quantity, classical nucleation theory¹ is commonly used to analyse crystal nucleation experiments, with the unconstrained parameters adjusted to fit the observations. This approach yields no ‘first principles’ prediction of absolute nucleation rates. Here we approach the problem from a different angle, simulating the nucleation process in a suspension of hard colloidal spheres, to obtain quantitative numerical predictions of the crystal nucleation rate. We find large discrepancies between the computed nucleation rates and those deduced from experiments^{2–4}: the best experimental estimates of P_{crit} seem to be too large by several orders of magnitude.

The probability (per particle) that a spontaneous fluctuation will

result in the formation of a critical nucleus depends exponentially on the free energy ΔG_{crit} that is required to form such a nucleus:

$$P_{\text{crit}} = \exp(-\Delta G_{\text{crit}}/k_{\text{B}}T) \quad (1)$$

where T is the absolute temperature and k_{B} is Boltzmann’s constant. According to classical nucleation theory (CNT), the total free energy of a crystallite that forms in a supersaturated solution contains two terms: the first is a ‘bulk’ term that expresses the fact that the solid is more stable than the supersaturated fluid—this term is negative and proportional to the volume of the crystallite. The second is a ‘surface’ term that takes into account the free-energy cost of creating a solid–liquid interface. This term is positive and proportional to the surface area of the crystallite. According to CNT, the total (Gibbs) free-energy cost to form a spherical crystallite with radius R is

$$\Delta G = \frac{4}{3}\pi R^3 \rho_{\text{s}} \Delta \mu + 4\pi R^2 \gamma \quad (2)$$

where ρ_{s} is the number-density of the solid, $\Delta \mu$ (<0) is the difference in chemical potential of the solid and the liquid, and γ is the solid–liquid interfacial free energy density. The function ΔG goes through a maximum at $R = 2\gamma/(\rho_{\text{s}}|\Delta \mu|)$ and the height of the nucleation barrier is:

$$\Delta G_{\text{crit}} = \frac{16\pi}{3} \gamma^3 / (\rho_{\text{s}} |\Delta \mu|)^2 \quad (3)$$

The crystal-nucleation rate per unit volume, I , is the product of P_{crit} and the kinetic prefactor κ :

$$I = \kappa \exp(-\Delta G_{\text{crit}}/k_{\text{B}}T) \quad (4)$$

The CNT expression for the nucleation rate then becomes:

$$I = \kappa \exp \left[-\frac{16\pi}{3} \gamma^3 / (\rho_{\text{s}} |\Delta \mu|)^2 \right] \quad (5)$$

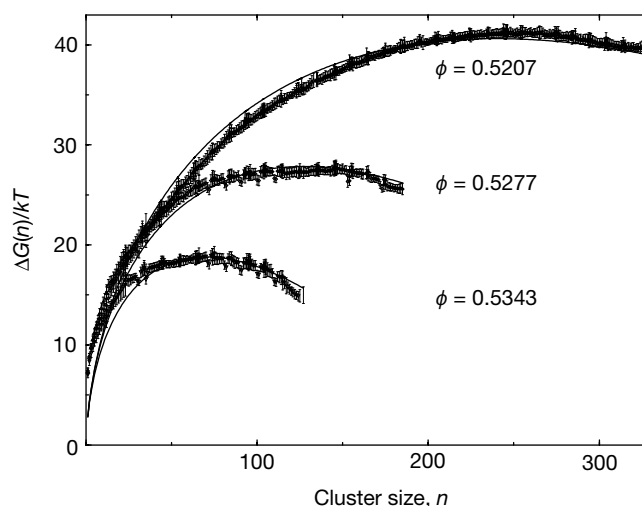


Figure 1 The calculated free-energy barrier for homogeneous crystal nucleation of hard-sphere colloids. Results are shown for three values of the volume fraction of the colloids in the liquid phase. $\phi_{\text{liq}} = 0.5207, 0.5277$ and 0.5343 . The barrier was computed using equation (6). A prerequisite for the calculation of the nucleation barrier is the choice of a ‘reaction coordinate’ that measures the progress from liquid to solid. As our reaction coordinate we use n , the number of particles that constitute the largest solid-like cluster in the system. A criterion based on that in ref. 8 was used to identify which particles are solid-like. If two solid-like particles are less than 2σ apart, where σ is the diameter of a particle, then they are counted as belonging to the same cluster. Using this technique we are able to distinguish between particles in a liquid-like environment and particles that belong to crystalline nuclei. For all but the smallest clusters, $P(n) \ll 1$. We used ‘umbrella sampling’²² to determine $P(n)$ in the range where it is very small. The total

simulation was split up into a number of smaller simulations that were restricted to a sequence of narrow, but overlapping ‘windows’ of n values. Stacking rearrangements in the crystalline nuclei were found to be slow. To alleviate this problem, we applied the parallel tempering scheme of Geyer and Thompson²³ to exchange clusters between adjacent windows. All simulations were performed at constant pressure and with the total number of particles (solid plus liquid) fixed. For every window, the simulations took at least 8×10^5 Monte Carlo moves per particle, excluding equilibration. The results of all simulations are presented in reduced units. In all cases, periodic boundary conditions were imposed. To eliminate noticeable finite-size effects, we simulated systems containing 3,375 hard spheres. The drawn curves are fits to the CNT expression (equation (2)). The fits yield the following values: $\gamma_{\text{eff}}(P = 15) = 0.699$, $\gamma_{\text{eff}}(P = 16) = 0.738$ and $\gamma_{\text{eff}}(P = 17) = 0.748$.

This expression has been used extensively to analyse crystal-nucleation experiments. However, as a rigorous theory for the kinetic prefactor is lacking, CNT has not been very successful in predicting absolute nucleation rates. Rather, the kinetic prefactor κ and, more often than not, the effective interfacial free energy γ are fitted to match the experimental nucleation rates¹. This situation is now changing. The computer simulations that we report here allow us to predict absolute crystal nucleation rates without making use of any adjustable fit parameters or, for that matter, of the assumptions underlying CNT. It is obviously important to compare the simulation results with crystal-nucleation experiments on a system that is very well characterized. For this reason, we chose to simulate crystal nucleation in suspensions of hard-sphere colloids. The hard-sphere freezing transition is probably better characterized than any other. Moreover, several groups have performed experimental studies of crystal nucleation in suspensions of hard-sphere colloids^{2–5}.

Experimental nucleation rate densities are usually expressed in dimensionless form: $I^* \equiv I\sigma^5/D_0$. Here σ is the hard-core diameter of the colloidal particles and D_0 is their self-diffusion coefficient at infinite dilution. In what follows, we always use reduced quantities and hence we will omit the asterisk. We use σ as our unit of length, $k_B T$ as our unit of energy and σ^2/D_0 as our unit of time. In the analysis of experiments on hard-core colloids, the kinetic prefactor κ is usually written as $A\phi_{\text{liq}}^{5/3}D(\phi_{\text{liq}})$, where ϕ_{liq} is the volume fraction of

the colloids in the liquid phase, $D(\phi)$ is the (known) reduced diffusivity at volume fraction ϕ , while A and γ are treated as adjustable parameters^{3,4,6}. We note that, even apart from the use of adjustable parameters to fit the data, the experimental tests of CNT for hard-sphere freezing are, at present, rather indirect as experiments do not probe the shape of the nucleation barrier directly. Nor do they provide information about the structure of the critical nucleus.

We have performed numerical simulations of hard-sphere colloids that allow us to compute the shape and height of the nucleation barrier and the structure of the critical nucleus. In addition, we have performed kinetic Monte Carlo simulations to compute the kinetic prefactor κ . These allow us to compare our simulation results for the reduced nucleation rate (without adjustable parameters) with the values of I determined in experiment^{2–4}.

To study the formation of a critical crystal nucleus, we used a biased Monte Carlo method^{7,8}. This scheme allows us to compute the equilibrium probability $P(n)$ for the formation of a crystalline cluster of size n . The (Gibbs) free energy of a cluster of size n is given by:

$$G(n) = \text{const.} - \ln[P(n)] \quad (6)$$

We computed the nucleation barrier as a function of the cluster size n for hard-sphere fluids that were compressed above the coexistence pressure $P_{\text{coex}} = 11.67$ (ref. 9). Simulations were performed at reduced pressures $P = 15, 16$ and 17 , corresponding to volume fractions of the liquid $\phi_{\text{liq}} = 0.5207, 0.5277$ and 0.5343 . These state points correspond to the lower range of supersaturations where hard-sphere nucleation has been studied experimentally^{2–4}. The reason for selecting this density regime is that at higher supersaturations, many crystal nuclei form simultaneously.

Figure 1 shows the computed excess Gibbs free energy associated with the formation of a cluster of size n in a supercooled liquid. The top of the barrier determines ΔG_{crit} and the critical nucleus size. Comparing our results for ΔG_{crit} with the best experimental estimates, we find that the latter are three times too low^{3,4}. One possible source of discrepancy could be that we simulated monodisperse suspensions, whereas the experimental systems have a size polydispersity of approximately 5%. We therefore repeated the simulations for a system with 5% polydispersity. However, we found that the only effect of polydispersity is to shift the coexistence curve. To within the numerical error in ΔG_{crit} ($\pm 1k_B T$), the monodisperse and polydisperse suspensions have the same nucleation

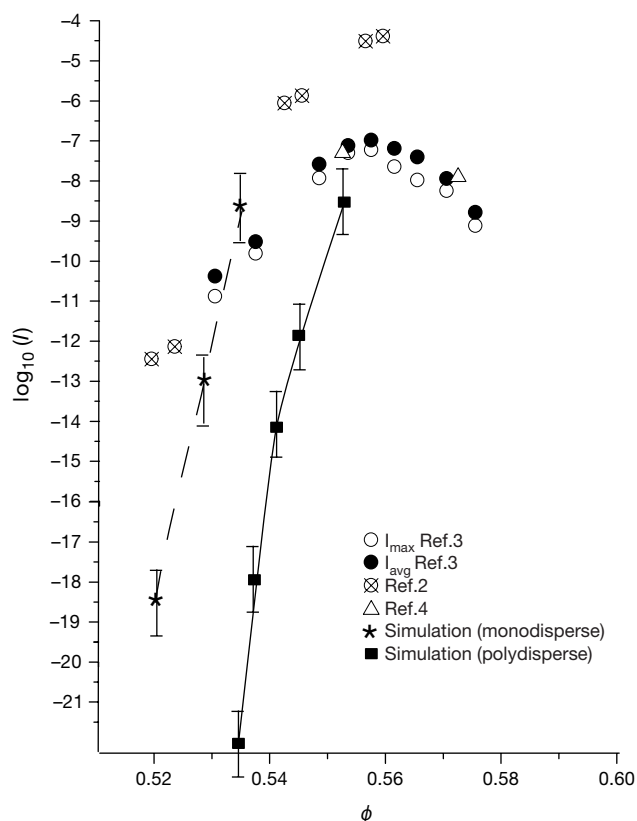


Figure 2 Reduced nucleation rates I as a function of the volume fraction of the metastable liquid. The simulation data for monodisperse colloids are indicated by asterisks—the drawn curve joining the simulation points is meant as a guide to the eye. We also show the experimental results of ref. 2 (crossed circles), ref. 3 (open and filled circles) and ref. 4 (triangles). We also performed simulations on model systems that have the same polydispersity (5%) as the experimental systems. These simulation results are denoted by filled squares. The discrepancy between the latter simulations and experiment is unexpected and significant. Several factors could complicate the comparison with the available experiments: first, the experiments yield a ‘time-averaged’ nucleation rate that may differ from the steady-state rate that we compute. Secondly, it is conceivable that the experimental systems contain some pre-critical nuclei.

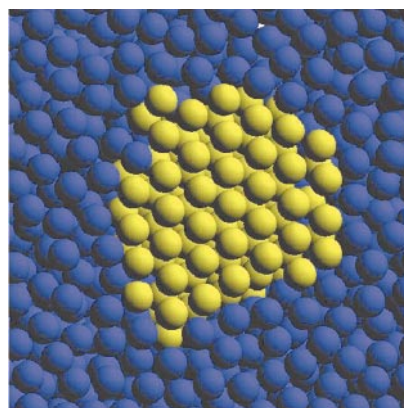


Figure 3 Snapshot of a cross-section of a critical nucleus of a hard-sphere crystal at a liquid volume fraction $\phi = 0.5207$. The figure shows a three-layer-thick slice through the centre of the crystallite. Solid-like particles are shown in yellow, and liquid-like particles in blue. The layers shown in the figure are close-packed hexagonal crystal planes. The stacking shown in this figure happens to be f.c.c.-like (that is, ABC stacking); however, analysis of many such snapshots showed that f.c.c. and h.c.p. stackings were equally likely.

barrier at the same supersaturation (that is, at the same value of $\Delta\mu$). We also compared the computed nucleation barriers with the predictions based on CNT. The dependence of ρ_s and $\Delta\mu$ on P can be accurately computed using the phenomenological equations of state for the solid and fluid phase of hard spheres¹⁰. For γ we take a recent numerical estimate¹¹. The resulting CNT predictions for ΔG_{crit} are 30–50% too low. These deviations are not small. For instance, for a barrier height of $40k_B T$, an error of 30% in ΔG_{crit} implies an error of the order of 10^5 in $\exp(-\Delta G_{\text{crit}}/k_B T)$ and the experimental estimate of $\exp(-\Delta G_{\text{crit}}/k_B T)$ is in error by a factor of the order of 10^{11} . Although the computed barrier heights are not predicted correctly by CNT, we can still fit our data to the functional form given by CNT (equation (2)). Using $n = 4\pi\rho_s R^3/3$, we can express equation (2) in terms of the cluster size n . The only adjustable parameter in our fit of the simulation data is the effective interfacial free-energy density γ_{eff} . Figure 1 shows that CNT reproduces the functional form of the nucleation barrier, except for very small clusters. The fit yielded the following values: $\gamma_{\text{eff}}(P = 15) = 0.699$, $\gamma_{\text{eff}}(P = 16) = 0.738$ and $\gamma_{\text{eff}}(P = 17) = 0.748$. Again, we find the same answer for the system with 5% polydispersity: that is, at the same supersaturation we obtain the same values for γ_{eff} of the monodisperse and polydisperse samples. We note that the numerical estimate of γ_{eff} is higher than the value deduced from the analysis of experimental nucleation data: the data analysis of refs 3 and 4 suggests that $\gamma_{\text{eff}} = 0.5$. As $\Delta G_{\text{crit}} \sim \gamma_{\text{eff}}^3$, the experimental estimate of ΔG_{crit} is a factor of three too low. If we assume that γ_{eff} depends linearly on pressure, then our simulation results extrapolate to a value of $\gamma \approx 0.62$ at coexistence—in good agreement with the numerical estimate¹¹. However, in the regime

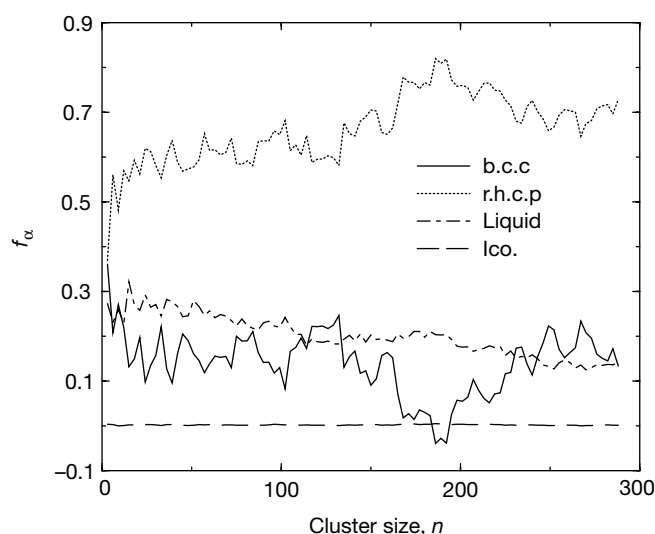


Figure 4 Structure analysis of (pre) critical crystal nuclei. The figure shows the relative weight of the structural signatures for r.h.c.p., b.c.c., icosahedral and liquid-like ordering in hard-sphere crystal nuclei of size n . In order to perform the structural analysis, we first computed the distribution of bond-order parameters for the various pure structures⁸. This is straightforward for liquid structures. In the case of the randomly stacked hexagonal close-packed structure, we determined the signature of a randomly stacked bulk crystal. However, no stable b.c.c. or icosahedral structures exist for monodisperse hard spheres. But we found that we could generate metastable b.c.c. structures for slightly (3%) polydisperse hard-sphere crystals. We computed the b.c.c. signature for this structure. In the case of the icosahedral structure, we computed the relevant bond-order parameter distributions for a particle that was constrained to be in an artificially stabilized icosahedral environment of the correct density. The figure shows that b.c.c. and icosahedral structures play no role in the nucleation process. Small clusters are fairly disordered and have an appreciable liquidlike signature. The figure also shows that the r.h.c.p. signature is dominant for all cluster sizes. However, small crystallites tend to be fairly disordered. In those structures, the b.c.c. and liquid-like signatures become noticeable.

where nucleation experiments have been performed, we find much larger values for γ_{eff} (between 0.7 and 0.75). CNT fails because it uses the value of γ at coexistence in equation (3) to predict ΔG_{crit} .

Our simulations allow us to give what we believe to be the first parameter-free estimate of the crystal nucleation rate. To this end, we need to compute the reduced kinetic prefactor κ . It has the following form: $\kappa = Z\rho f_{n_c}/D_0$, where Z is the Zeldovitch factor¹, ρ is the number density of the supersaturated liquid, and f_{n_c} is the addition rate of particles to the critical nucleus. The Zeldovitch factor $Z = [|\Delta G''(n_{\text{crit}})|/(2\pi k_B T)]^{1/2}$, where $\Delta G''(n)$ is the second derivative of $\Delta G(n)$, can be obtained directly from the simulation results. We have computed f_{n_c}/D_0 using kinetic Monte Carlo simulations of the hard-sphere suspension¹². In such simulations, the effect of hydrodynamic interactions between colloids is ignored. However, in line with ref. 13, we correct for this effect by multiplying our Monte Carlo results for f_{n_c} by a factor $\alpha(\phi) \equiv D^s(\phi)/D_0$, where $D^s(\phi)$ is the short-time self-diffusion coefficient at volume fraction ϕ . Several, rather similar, functional forms for $\alpha(\phi)$ have been proposed in the literature. Here we use the phenomenological expression $\alpha(\phi) = (1 - \phi/0.64)^{1.17}$ (ref. 6). When applying the same approach to the computation of the long-time self-diffusion constant, we reproduce the experimental data in the same density range (see ref. 4) to within the statistical error. We estimate that the error in $\ln \kappa$ is ± 1 , that is, about the same as the error in $\Delta G_{\text{crit}}/k_B T$.

Figure 2 shows our numerical estimate for the reduced nucleation rate I . We have computed I both for a monodisperse suspension and for a suspension with 5% polydispersity. The latter results can be compared directly with the experimental studies of steady-state nucleation^{2–4}. As can be seen in Fig. 2, the simulated and experimental curves almost intersect, but for most densities we observe discrepancies of many orders of magnitude. Such discrepancies are both real and a cause for concern, as we estimate that our computed nucleation rates are accurate to within one order of magnitude. We therefore argue that the problem must be with the interpretation of the experiments, and hope that future ‘real-space’ experiments will provide an explanation for the discrepancy in the computed nucleation rates and barrier heights.

The simulations allow us to address a question that cannot, at present, be answered experimentally: what is the structure of the critical nucleus? Figure 3 shows an example of a snapshot of a critical nucleus observed in our simulations. CNT makes the assumption that the (pre) critical nuclei are effectively spherical, and have the same structure as the stable bulk phase that is nucleating. However, Ostwald¹⁴ pointed out in 1897 that the phase that nucleates need not be the one that is thermodynamically stable. In recent years, several attempts have been made to provide a microscopic explanation for Ostwald’s observation^{15–17}. Alexander and McTague¹⁵ have argued, on the basis of Landau theory, that in the early stages of crystal nucleation a body-centred cubic (b.c.c.) crystallite should form that would subsequently transform into the stable crystal phase. Indeed, simulations of pre-critical nuclei in a Lennard–Jones (‘argon’) liquid are found to have b.c.c. structure, rather than the stable face-centred cubic (f.c.c.)—but this f.c.c. phase is only slightly more stable than the hexagonal close packed (h.c.p.) phase¹⁸. In fact, the f.c.c.–h.c.p. free-energy difference is so small that small hard-sphere crystals will always contain an equilibrium concentration of stacking faults^{5,19–21}. It would clearly be interesting to know by what route hard-sphere crystals nucleate: does the small size lift the effective degeneracy between f.c.c. and h.c.p. structures to the extent that one of these dominates, or do hard-sphere crystal nuclei obey the predictions of the Alexander-McTague theory and exhibit a b.c.c. or possibly even icosahedral structure?

To answer this question we analysed the structure of the crystalline nucleus. The stable structure of a bulk hard-sphere solid is f.c.c.¹⁸. But as the free energy of stacking faults is small^{20,21}, small crystallites are expected to exhibit random stacking of f.c.c. and

h.c.p. domains. Indeed, direct inspection of the crystal nuclei generated in our simulations show that the critical nucleus is randomly stacked. The next question to ask is if there is any evidence for b.c.c. or icosahedral ordering. To this end, we used the analysis technique of ten Wolde *et al.*⁸. This approach allows us to analyse any cluster structure as a mixture of simple reference structures. We assume that pre-critical nuclei need not have the random-hexagonal close-packed (r.h.c.p.) structure. Rather, we allow for the possibility that the nuclei exhibit a signature characteristic of b.c.c., icosahedral or even liquid-like ordering. Every structure is characterized by a set of numbers $\{f_{bcc}, f_{rhcp}, f_{icos}, f_{liq}\}$, where the value of f_α denotes the relative importance of structure α in the cluster. In Fig. 4 we show the results for f_{bcc} , f_{rhcp} , f_{icos} and f_{liq} as a function of the size of the largest cluster in the system at $P = 15$. The results for $P = 16$ and 17 are qualitatively similar. The figure shows that icosahedral ordering is not observed for any nucleus size. Small clusters still have some b.c.c. or liquid-like signature. But in all cases the r.h.c.p. signature is dominant. The same conclusion holds for the weakly polydisperse crystals that we studied. This observation shows that the simplest of all crystals does not confirm the theoretical prediction that b.c.c. (or icosahedral) (pre)nuclei should be favoured in the crystallization of simple liquids^{15–17}. Our finding is also unexpected in the light of the finding that pre-critical Lennard–Jones nuclei have a b.c.c. structure⁸. As the structure of simple liquids is known to be dominated by the short-ranged repulsive forces, we expected that the structure of the pre-critical nucleus in hard-sphere fluids would be b.c.c., as in the Lennard–Jones case.

Random hexagonal close packing (r.h.c.p. stacking) in freshly nucleated colloidal hard-sphere crystals has been observed in several experiments^{5,19}. It is usually assumed that the origin of the r.h.c.p. stacking is purely kinetic. Although this may be correct for larger crystals, the present simulation indicates that, in the early stages of nucleation, the r.h.c.p. structure is simply more stable than f.c.c. This phenomenon can be interpreted as a manifestation of Ostwald's 'step rule'¹⁴. The hard-sphere fluid nucleates into the metastable r.h.c.p. structure. Only later does this metastable structure transform into the stable f.c.c. structure^{20,21}. □

Received 21 July; accepted 29 November 2000.

- Kelton, K. F. *Solid State Physics* Vol. 45, 75–90 (Academic, New York, 1991).
- Schätzel, K. & Ackerson, B. J. Density fluctuations during crystallization of colloids. *Phys. Rev. E* **48**, 3766–3777 (1993).
- Harland, J. L. & Van Megen, W. Crystallization kinetics of suspensions of hard colloidal spheres. *Phys. Rev. E* **55**, 3054–3067 (1997).
- Cheng, Z. *Colloidal Hard Sphere Crystallization and Glass Transition*. Thesis, Princeton Univ. (1998).
- Zhu, J. *et al.* Crystallization of hard-sphere colloids under microgravity. *Nature* **387**, 883–885 (1997).
- Van Duijneldt, J. S. & Lekkerkerker, H. N. W. in *Science and Technology of Crystal Growth* (eds van der Eerden, J. P. & Bruinsma, O. S. L.) 279–290 (Kluwer Academic, Dordrecht, 1995).
- Van Duijneldt, J. S. & Frenkel, D. Computer simulation study of free energy barriers in crystal nucleation. *J. Chem. Phys.* **96**, 4655–4668 (1992).
- Ten Wolde, P. R., Ruiz-Montero, M. J. & Frenkel, D. Numerical evidence for b.c.c. or ordering at the surface of a critical f.c.c. nucleus. *Phys. Rev. Lett.* **75**, 2714–2717 (1995).
- Hoover, W. G. & Ree, F. H. Melting transition and communal entropy for hard spheres. *J. Chem. Phys.* **49**, 3609–3617 (1968).
- Hall, K. R. Another hard-sphere equation of state. *J. Chem. Phys.* **57**, 2252–2254 (1970).
- Davidchack, R. L. & Laird, B. B. Direct calculation of the hard-sphere crystal/melt interfacial free energy. *Phys. Rev. Lett.* **85**, 4751–4754 (2000).
- Hinsen, K. & Cichocki, B. Dynamic computer simulation of concentrated hard-sphere suspensions. *Physica A* **166**, 473–491 (1990).
- Medina-Noyola, M. Long-time self-diffusion in concentrated colloidal dispersions. *Phys. Rev. Lett.* **60**, 2705–2708 (1988).
- Ostwald, W. Studien über die Bildung und Umwandlung fester Körper. *Z. Phys. Chem.* **22**, 289–330 (1897).
- Alexander, S. & McTague, J. P. Should all crystals be BCC? *Phys. Rev. Lett.* **41**, 702–705 (1978).
- Klein, W. & Leyvraz, F. Crystalline nucleation in deeply quenched liquids. *Phys. Rev. Lett.* **57**, 2845–2848 (1986).
- Groh, B. & Mulder, B. M. Why all crystals need not be b.c.c.: symmetry breaking at the liquid–solid transition revisited. *Phys. Rev. E* **59**, 5613–5620 (1999).
- Bolhuis, P. G., Frenkel, D., Mau, S. C. & Huse, D. A. Entropy difference between the face-centred cubic and hexagonal close-packed crystal structures. *Nature* **388**, 235–237 (1997).
- Pusey, P. N. *et al.* Structure of crystals of hard colloidal spheres. *Phys. Rev. Lett.* **63**, 2753–2756 (1989).
- Pronk, S. & Frenkel, D. Can stacking faults in hard-sphere crystals anneal out spontaneously? *J. Chem. Phys.* **110**, 4589–4592 (1999).
- Mau, S. C. & Huse, D. A. Stacking entropy of hard-sphere crystals. *Phys. Rev. E* **59**, 4396–4401 (1999).

- Torrie, G. M. & Valleau, J. P. Monte-Carlo free energy estimates using non-Boltzmann sampling: application to the subcritical Lennard–Jones fluid. *Chem. Phys. Lett.* **28**, 578–581 (1974).
- Geyer, C. J. & Thompson, E. A. Annealing Markov-chain Monte Carlo with applications to ancestral inference. *J. Am. Stat. Assoc.* **90**, 909–920 (1995).

Acknowledgements

We thank H. Lekkerkerker, W. van Megen, W. Kegel, A. van Blaaderen, J. Horbach and B. Smit for a critical reading of the manuscript. This work was supported by the Division of Chemical Sciences (CW) of the Netherlands organization for Scientific Research (NWO). The work of the FOM Institute is part of the research programme of FOM, and is made possible by financial support from the NWO.

Correspondence and requests for materials should be addressed to D.F. (e-mail: frenkel@amolf.nl).

Formation of thermally stable alkylidene layers on a catalytically active surface

El Mamoune Zahidi, Hicham Oudghiri-Hassani & Peter H. McBreen

Département de Chimie et CERPIC, Université Laval, Québec (PQ), G1K 7P4, Canada

Materials containing organic–inorganic interfaces usually display a combination of molecular and solid-state properties, which are of interest for applications ranging from chemical sensing¹ to microelectronics² and catalysis³. Thiols—organic compounds carrying a SH group—are widely used to anchor organic layers to gold surfaces⁶, because gold is catalytically sufficiently active to replace relatively weak S–H bonds with Au–S bonds, yet too inert to attack C–C and C–H bonds in the organic layer. But although several methods^{4–6} of functionalizing the surfaces of semiconductors, oxides and metals are known, it remains difficult to attach a wide range of more complex organic species. Organic layers could, in principle, be formed on the surfaces of metals that are capable of inserting into strong bonds, but such surfaces catalyse the decomposition of organic layers at temperatures above 400 to 600 K, through progressive C–H and C–C bond breaking⁷. Here we report that cycloketones adsorbed on molybdenum carbide, a material known to catalyse a variety of hydrocarbon conversion reactions^{8–11}, transform into surface-bound alkylidenes stable to above 900 K. We expect that this chemistry can be used to create a wide range of exceptionally stable organic layers on molybdenum carbide.

The experiments were performed by exposing clean molybdenum carbide, in the form of β -Mo₂C, to cycloketones and then probing the surface using infrared reflectance, mass spectrometric thermal desorption, X-ray photoemission (XPS) and chemical reactivity methods. We observed two intense desorption peaks appearing between 900 and 1,200 K, shown in Fig. 1. This feature is common to the data for C₄–C₆ cycloketones, and mass spectrometric calibration data (see Supplementary Information) identify the desorbing species, in each case, as the parent cycloketone molecule. These high-temperature peaks must then arise from ketone-formation reactions, because intact small molecules typically desorb below 400 K. Furthermore, the intermediates for ketone formation must be stable to above 900 K, in contrast to previous observations⁷ of the stability of complex organic species on materials as catalytically active^{8–11} as β -Mo₂C. Infrared reflectance spectra (Fig. 2a) of a cyclobutanone monolayer adsorbed on β -Mo₂C at 105 K display a carbonyl stretching vibration band at 1,730 cm^{–1} characteristic of a chemisorption bond between the carbonyl oxygen and the surface.

On heating to 300 K, the carbonyl band disappears and a Mo=O (oxo) band¹² appears at 950 cm⁻¹. The $\nu_{as}(\text{CH}_2)$ band at 2,967 cm⁻¹ (Fig. 2b) is characteristic^{13,14} of highly strained cyclic molecules, and the band at 1,130 cm⁻¹ is close in frequency to the value of 1,121 cm⁻¹ assigned to the metal–carbon vibration of Ti-neopentylidene¹⁵. These observations are consistent with carbonyl bond scission in our system, to generate a surface alkylidene–Mo-oxo complex (Fig. 3), in direct analogy to the oxidative addition of cyclopentanone to $\text{WCl}_2(\text{PMePh}_2)_4$ to form a cyclopentylidene–W-oxo product^{16,17}. Isotope exchange results (see Supplementary Information) show incorporation of ¹⁸O into the species desorbing at high temperatures, presumably through a process that is the reverse of the carbonyl bond scission reaction occurring at low temperatures. Cyclobutanone possesses a strain energy of about ~29.0 kcal mol⁻¹, and its carbon ring is therefore unlikely simply to open and close again during the course of our experiments. This implies that the intermediate species, formed upon cyclobutanone adsorption and from which cyclobutanone materializes again at high temperatures, contains the C₄ ring. This is confirmed by the infrared data in Fig. 2b, which provides evidence for the presence of cyclobutylidene at 900 K. Direct chemical proof for the presence of the cyclobutylidene Mo=C double bond is given by the results of olefin metathesis test reactions, carried out by exposing the surface, after pre-treatment with cyclobutanone, to propene and then recording thermal desorption spectra. The products detected (Fig. 4) are methylene cyclobutane at low temperatures and acetaldehyde at high temperatures, as would be expected for olefin metathesis proceeding through the accepted Hérisson–Chauvin metallacyclobutane mechanism^{18,19}. In this mechanism, the alkylidene and olefin double bonds add to form a four-membered ring containing the metal atom; this intermediate

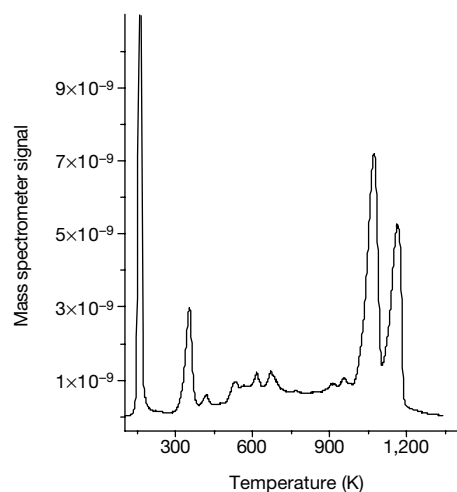


Figure 1 High-temperature formation of cyclobutanone on the catalytically active $\beta\text{-Mo}_2\text{C}$ surface. The carbide sample, displaying the bulk hexagonal structure, was prepared by temperature ramping pure molybdenum foil in a CH_4/H_2 mixture²⁰. The surface was cleaned using a combination of ion sputtering to remove sulphur, treatment in propylene at 500 K to remove oxygen, and annealing to 1,400 K to drive excess carbon into the subsurface. All experiments were performed, under ultrahigh-vacuum conditions (base pressure equal to 7×10^{-11} Torr), using samples displaying a stoichiometric surface region as determined from XPS measurements. The temperature-programmed desorption spectrum, displaying the mass to charge ratio $m/e=42$ data, was obtained by exposing the carbide at 105 K to 120×10^{-6} Torr s cyclobutanone and then heating the sample at 1 K s⁻¹. The sharp low-temperature feature is due to the desorption of condensed, multilayer cyclobutanone. Most of the remaining monolayer dissociates into fragments forming strong bonds to the surface. The high-temperature double peak is due to a surface reaction in which fragments recombine to regenerate the parent molecule. Exposures of approximately 60×10^{-6} Torr s are required to saturate the high-temperature desorption signal.

then undergoes simultaneous cleavage of opposite sides of the metallacycle to yield a new olefin and a new alkylidene. In the case of propene and Mo-cyclobutylidene, the new olefin is methylene cyclobutane, which desorbs at relatively low temperatures, and the new alkylidene is CH_3CHMo , from which acetaldehyde is formed by reaction with surface oxygen.

The preparation of ultrastable $\beta\text{-Mo}_2\text{C}$ -alkylidenes using cycloketones depends on a subtle balancing of surface reactivities: the surface must initially be reactive enough to break the very strong carbonyl bond at low temperatures, but not so reactive that it decomposes the resultant surface cycloalkylidene at higher temperatures. H_2 desorption measurements, as a function of cycloketone exposure, show that these competing requirements are spontaneously met, while the surface is progressively passivated as a result of ketone decomposition at low exposures. Non-selective decomposition of the ketone gives an H_2 desorption peak in a broad mid-temperature range, and carbon deposition, determined by XPS measurements, as expected for hydrocarbons on catalytically active surfaces⁷. However, as the exposure is increased, the relative amount of hydrogen desorption drops markedly and selective high-temperature desorption of the ketone is observed. Studies using isotopically labelled cyclohexanone mixtures show that the condensed layer prepared at 105 K participates in the high-temperature

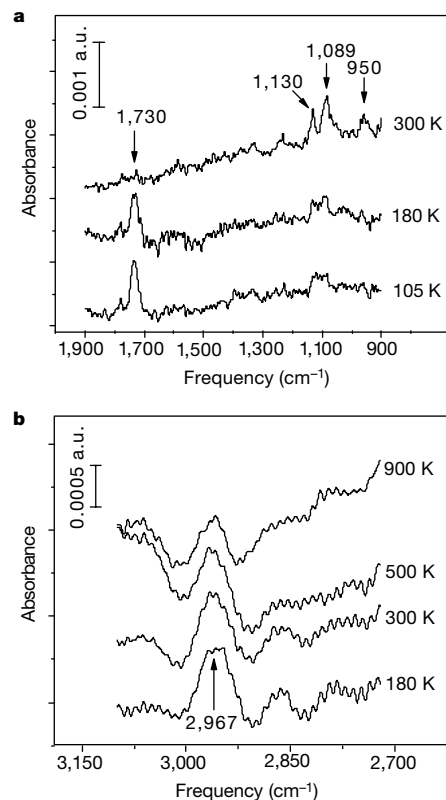


Figure 2 Infrared reflectance spectra showing the cleavage of the cyclobutanone carbonyl bond and the formation of surface cyclobutylidene. The spectra are expressed in absorbance units (a.u.) and correspond to the ratio of 800 sample scans to 800 background scans of the clean sample. All measurements were made at 105 K and approximately 300 s were required to cool the sample from 900 to 105 K. Exposures above 3×10^{-6} Torr s lead to the appearance of bands due to condensed cyclobutanone. **a**, The absorption bands at 950, 1130 and 1730 cm⁻¹ are characteristic of oxo (Mo=O)¹², metal–alkylidene (Mo=C)¹⁵, and carbonyl (C=O) bonds, respectively. **b**, The absorption band at 2,967 cm⁻¹ is characteristic of the CH_2 asymmetric stretching vibration ν_{as} in highly strained rings such as cyclobutanone or cyclobutylidene^{13,14}. The persistence of the $\nu_{as}(\text{CH}_2)$ band to 900 K shows the exceptional thermal stability of the surface alkylidene. The spectra in **b** were obtained following an exposure of 120×10^{-6} Torr s.

ketone-formation reaction, and also show that saturation of the thermally stable layer requires initial exposures greater than a monolayer. The structure of the high-temperature double desorption peak (Fig. 1) is very similar to that reported for H_2O formation in the reduction of molybdenum and tungsten oxides in CH_4/H_2

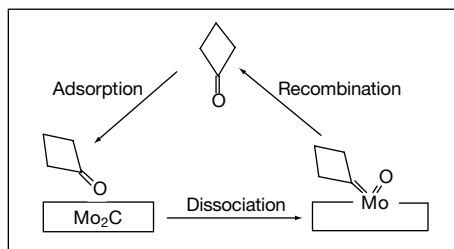


Figure 3 Schematic illustration of reversible cleavage of the cycloketone carbonyl bond mediated by alkylidene-Mo-oxo. The reactive surface breaks the very strong carbonyl bond at low temperatures. The resultant alkylidene remains attached to the surface until atomic oxygen inserts into the metal–carbon bond at high temperatures, leading to immediate desorption of the cycloketone product. The oxygen insertion reaction is confirmed by isotope exchange measurements (see Supplementary Information) showing that the interaction of cyclobutanone ($\text{C}_4\text{H}_6\text{O}$) with $^{18}\text{O}_2$ exposed $\beta\text{-Mo}_2\text{C}$ leads to the high temperature desorption of $\text{C}_4\text{H}_6^{18}\text{O}$. The illustration is not meant to imply that the oxo state remains stable between 300 and 900 K.

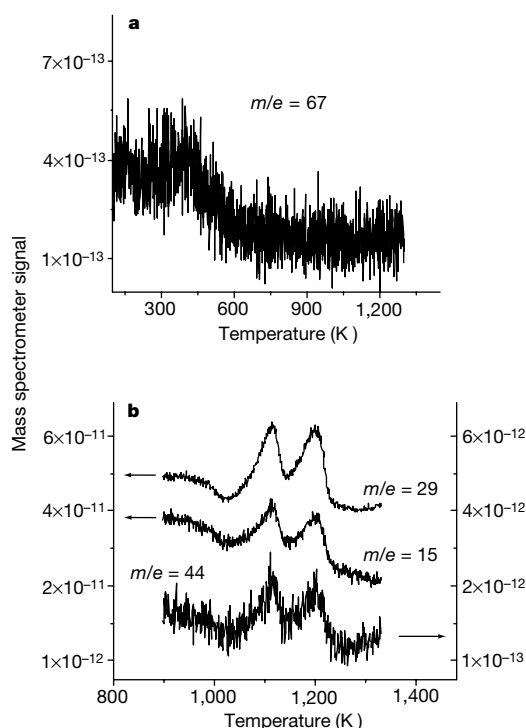


Figure 4 Direct chemical determination of the presence of surface alkylidenes by performing an olefin metathesis reaction. The test reaction was carried out by exposing $\beta\text{-Mo}_2\text{C}$ to 20×10^{-6} Torr s cyclobutanone at 105 K, heating to 200 K and then exposing the surface to 120×10^{-6} Torr s propene at 105 K. The observed products may be predicted on the basis of the Hérisson–Chauvin metallacyclobutane mechanism¹⁸ for olefin metathesis. Productive cleavage of the metallacycle intermediate formed through the [2+2] cycloaddition of the propene and cyclobutylidene double bonds leads to the addition of the propene CH_2 group to the surface alkylidene thereby forming the volatile methylene cyclobutane product (a). Simultaneously, the CHCH_3 moiety from propene forms a surface alkylidene group. The latter group is then eliminated from the surface as acetaldehyde (b) in the same type of oxygen-insertion reaction as that giving rise to the high-temperature regeneration of ketone. Metathesis activity was also observed for surfaces pre-treated to 500 and 900 K.

mixtures^{20,21}. Cycloketone regeneration may thus be viewed as the sequential removal of oxygen from the carbide in two final steps. Experiments using $^{18}\text{O}_2$ reveal that surface oxygen removes excess carbon from the surface, as CO, just below the ketone-formation temperature. We presume that this cleaning of the surface triggers the microscopic reversal of the carbonyl scission reaction seen at low temperatures, leading to recombination and desorption at high temperatures.

Thermally stable hydrocarbon species containing sp^2 carbons have been observed in the formation of silicon carbide layers on silicon²². It may be that carbides, in general, are effective substrates for the preparation of thermally stable hydrocarbon layers. The present result is, however, unprecedented in that it shows that thermal stability is compatible with a catalytically active surface. Furthermore, it introduces a new method for the preparation of complex and functionally active surface groups. Given the vast range of cyclic ketones and carbonyl containing molecules available to the chemist, the formation of exceptionally stable alkylidenes on molybdenum carbide should make the procedure useful for creating targeted organic layers. Early transition metal carbides are known to function as natural electrical contacts to carbon nanotubes²³ and their functionalization—particularly with multiple bonds²⁴, as in the reported reaction—will be an additional tool for the development of molecular electronics materials. The unprecedented thermal stability of the surface alkylidenes prepared in this study means that they may be used as spacer groups for high-temperature applications, or as chiral modifiers for the purposes of asymmetric heterogeneous catalysis. Molybdenum carbide is an excellent catalyst for several processes, especially since it forms naturally from Mo and MoO_3 during the induction period of several hydrocarbon conversion reactions^{25,26}. In particular, the olefin metathesis activity of alkylidene-functionalized molybdenum carbide points to its use as a template for the directed growth of polymer structures^{3,27,28}. □

Received 21 August 2000; accepted 19 January 2001.

- Crooks, R. M. & Ricco, A. J. New organic materials suitable for use in chemical sensor arrays. *Acc. Chem. Res.* **31**, 219–227 (1998).
- Wolkow, R. A. Controlled molecular adsorption on silicon: laying a foundation for molecular devices. *Annu. Rev. Phys. Chem.* **50**, 413–441 (1999).
- Jeon, N. L. *et al.* Patterned polymer growth on silicon surfaces using microcontact printing and surface-initiated polymerization. *Appl. Phys. Lett.* **75**, 4201–4202 (1999).
- Buriak, J. M. Organometallic chemistry on silicon surfaces: formation of functional monolayers bound through Si–C bonds. *Chem. Commun.* 1051–1060 (1999).
- Yan, C., Zharnikov, M., Götzhäuser, A. & Grunze, M. Preparation and characterization of self-assembled monolayers on indium tin oxide. *Langmuir* **16**, 6208–6215 (2000).
- Bain, C. D. *et al.* Formation of monolayer films by the spontaneous assembly of organic thiols from solution onto gold. *J. Am. Chem. Soc.* **111**, 321–335 (1989).
- Somorjai, G. A. *Introduction to Surface Chemistry and Catalysis* 420 (John Wiley & Sons, New York, 1994).
- Oyama, S. T. in *The Chemistry of Transition Metal Carbides and Nitrides* (Blackie, London, 1996).
- York, A. P. E. *et al.* Molybdenum and tungsten carbides as catalysts for the conversion of methane to synthesis gas using stoichiometric feedstocks. *Chem. Commun.* 39–40 (1997).
- Chen, J. G., Fröhberger, B., Eng Jr, J. & Bent, B. E. Controlling surface reactivities of transition metals by carbide formation. *J. Mol. Catal. A* **131**, 285–299 (1998).
- Solymosi, F. & Szöke, A. Conversion of ethane into benzene on an $\text{Mo}_2\text{C}/\text{ZSM-5}$ catalyst. *Appl. Catal. A* **166**, 225–235 (1998).
- Nugent, W. A. & Mayer, J. M. *Metal-Ligand Multiple Bonds* Ch. 4 (Wiley, New York, 1988).
- Silverstein, R. M., Bassler, G. C. & Morrill, T. C. *Spectrometric Identification of Organic Compounds* 5th edn, Ch. 3 (Wiley, New York, 1991).
- Cataliotti, R., Giorgini, M. G., Paliani, G. & Poletti, A. The vibrational spectrum of crystalline cyclobutanone. *Spectrochim. Acta A* **31**, 1879–1884 (1975).
- Cheon, J., Dubois, L. H. & Girolami, G. S. Mechanistic studies of the thermolysis of tetraeneopentyltitanium(IV). 2. Solid state and ultra-high-vacuum studies of the chemical vapor deposition of TiC films. *J. Am. Chem. Soc.* **119**, 6814–6820 (1997).
- Bryan, J. C. & Mayer, J. M. Oxidative addition of cyclopentanone to $\text{WCl}_2(\text{PMePh}_2)_4$ to give a tungsten(VI) oxo-alkylidene complex. *J. Am. Chem. Soc.* **109**, 7213–7214 (1987).
- Bryan, J. C. & Mayer, J. M. Oxidative addition of carbon-oxygen and carbon-nitrogen double bonds to $\text{WCl}_2(\text{PMePh}_2)_4$. Synthesis of tungsten metalloxirane and tungsten oxo- and imido-alkylidene complexes. *J. Am. Chem. Soc.* **112**, 2298–2308 (1990).
- Spessard, G. O. & Miessler, G. L. *Organometallic Chemistry* 348 (Prentice-Hall, New Jersey, 1997).
- Furstner, A. Olefin metathesis and beyond. *Angew. Chem. Int. Edn Engl.* **39**, 3012–3043 (2000).
- Kapoor, R. & Oyama, S. T. in *Synthesis and Characterization of Advanced Materials* (eds Gruen, D. M. & Malhotra, R.) Ch. 18 (ACS symposium series, American Chemical Society, Washington, 1998).

21. Lofberg, A., Frennet, A., Leclercq, G., Leclercq, L. & Giraudon, J. M. Mechanism of WO_3 reduction and carburization in CH_4/H_2 mixtures leading to bulk tungsten carbide powder catalysts. *J. Catal.* **189**, 189–193 (2000).
22. Park, S.-C., Park, H. & Lee, S. B. Reaction intermediate in thermal decomposition of 1,3-disilabutane to silicon carbide on Si(111). Comparative study of Cs^+ reactive ion scattering and secondary ion mass spectrometry. *Surf. Sci.* **450**, 117–125 (2000).
23. Zhang, Y. *et al.* Heterostructures of single-walled carbon nanotubes and carbide nanorods. *Science* **285**, 1719–1722 (1999).
24. Mayr, A., Yu, M. P. Y. & Yam, V. W.-W. Electronic communication between metal centers across unsaturated alkylidyne ligands. *J. Am. Chem. Soc.* **121**, 1760–1761 (1999).
25. Wang, D., Lunsford, J. H. & Rosynek, M. P. Characterization of a Mo/ZSM-5 catalyst for the conversion of methane to benzene. *J. Catal.* **169**, 347–358 (1997).
26. Tanaka, K.-I. & Takeshiro, N. Atomic-scale mechanism for the activation of catalyst surfaces. *J. Mol. Catal. A* **141**, 39–55 (1999).
27. Weck, M., Jackiw, J. W., Rossi, R. R., Weiss, P. S. & Grubbs, R. H. Ring-opening metathesis polymerization from surfaces. *J. Am. Chem. Soc.* **121**, 4088–4089 (1999).
28. Watson, K. J., Zhu, J., Nguyen, S. T. & Mirkin, C. A. Hybrid nanoparticles with block copolymer shell structures. *J. Am. Chem. Soc.* **121**, 462–463 (1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We acknowledge the technical assistance of A. Bouffard and J. Lafrèrère, and financial support from the National Science and Engineering Research Council (NSERC) and Le Fonds pour la Formation de Chercheurs et l'Aide à la Recherche (FCAR). We thank S. T. Oyama for the preparation of the molybdenum carbide samples.

Correspondence and requests for materials should be addressed to P.M.C.B. (e-mail: peter.mcbreen@chm.ulaval.ca).

Recent mass balance of polar ice sheets inferred from patterns of global sea-level change

Jerry X. Mitrovica*, Mark E. Tamisiea*, James L. Davis† & Glenn A. Milne‡

* Department of Physics, University of Toronto, 60 St George Street, Toronto, M5S 1A7, Canada

† Harvard-Smithsonian Center for Astrophysics, 60 Garden Street MS42, Cambridge, Massachusetts 02138, USA

‡ Department of Geological Sciences, University of Durham, Science Laboratories, South Road, Durham DH1 3LE, UK

Global sea level is an indicator of climate change^{1–3}, as it is sensitive to both thermal expansion of the oceans and a reduction of land-based glaciers. Global sea-level rise has been estimated by correcting observations from tide gauges for glacial isostatic adjustment—the continuing sea-level response due to melting of Late Pleistocene ice—and by computing the global mean of these residual trends^{4–9}. In such analyses, spatial patterns of sea-level rise are assumed to be signals that will average out over geographically distributed tide-gauge data. But a long history of modelling studies^{10–12} has demonstrated that non-uniform—that is, non-eustatic—sea-level redistributions can be produced by variations in the volume of the polar ice sheets. Here we present numerical predictions of gravitationally consistent patterns of sea-level change following variations in either the Antarctic or Greenland ice sheets or the melting of a suite of small mountain glaciers. These predictions are characterized by geometrically distinct patterns that reconcile spatial variations in previously published sea-level records. Under the—albeit coarse—assumption of a globally uniform thermal expansion of the oceans, our approach suggests melting of the Greenland ice complex over the last century equivalent to $\sim 0.6 \text{ mm yr}^{-1}$ of sea-level rise.

Gravitationally self-consistent sea-level changes arising from the growth or ablation of ice masses have been of interest for more than

a century^{10–12}. Woodward¹⁰ demonstrated that the melting of an ice mass on a rigid Earth would lead to a highly non-uniform sea-level redistribution as a consequence of self-gravitation in the surface load. Indeed, sea level on a rigid Earth will drop within $\sim 20^\circ$ of a localized (point mass) ice melting event¹². The sea-level theory was extended to include elastic deformations of the solid Earth (ref. 11 and others), culminating in the 'sea-level equation' derived and solved by Farrell and Clark¹².

We have computed sea-level redistributions associated with present-day mass variations in the Antarctic and Greenland ice complexes as well as melting from a suite of smaller land-based ice sheets and glaciers tabulated by Meier¹³. Our calculations are based on a new sea-level theory¹⁴ that extends earlier work¹² to include a varying shoreline geometry and the influence of load-induced perturbations in the Earth's rotation vector. As we are concerned with sea-level variations associated with relatively rapid ice flux scenarios, we adopt a form of the theory suitable for an elastic Earth. The elastic and density structure of the model are adopted from PREM¹⁵.

Meier's sources¹³ combine to provide a 'eustatic' sea-level rise of

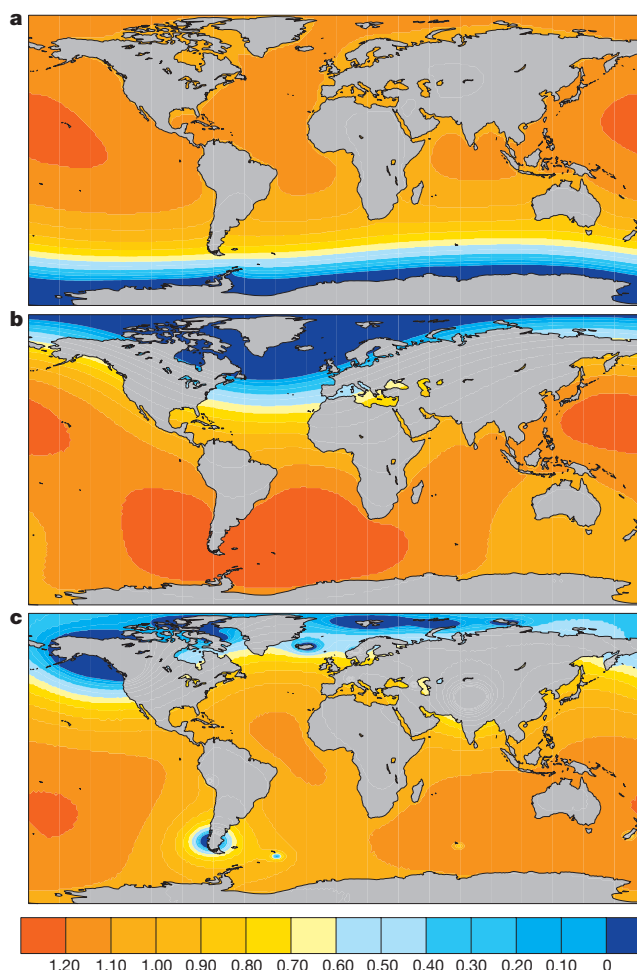


Figure 1 Predicted geometries of sea-level change due to continuing ice mass variations. Normalized global sea-level variations were computed for the case of present-day ice mass variations in **a**, Antarctica, **b**, Greenland and **c**, melting of the mountain glaciers and ice sheets tabulated by Meier¹³. In **a** and **b** we assume that the mass variation is uniform over the two polar regions. The results are normalized by the equivalent eustatic sea-level change for each mass flux event (see text). Departures from a contour value of 1.0 reflect departures from the assumption that the sea-level distribution accompanying these mass flux events is uniform. Predictions are based on a new sea-level theory¹⁴ solved using a pseudo-spectral algorithm^{14,29} with truncation at spherical harmonic degree and order 512. This truncation corresponds to a spatial resolution of 40 km.

0.46 mm yr⁻¹. (In this context, the term eustatic is defined as the change in ocean volume, that is, the change in ice mass divided by the density of water, divided by the area of the ocean.) The integrated mass fluxes from the Antarctic and Greenland ice sheets are quantities of intense interest and they remain uncertain, even in sign^{1,16}. Accordingly, we simply assume a uniform ice mass variation over the entire ice sheet in each case.

Figure 1 shows global maps of sea-level change due to mass variations in the Antarctic, Greenland and Meier's sources¹³. The frames in Fig. 1 are normalized by the eustatic sea-level variation associated with each of the three ice systems, which we denote by V_A , V_G , and V_M , respectively. We can thus write:

$$S_i(\theta, \phi) = V_A \times S_A''(\theta, \phi) + V_G \times S_G''(\theta, \phi) + V_M \times S_M''(\theta, \phi) \quad (1)$$

where $S''(\theta, \phi)$ are the maps of normalized sea-level change in Fig. 1, and $S_i(\theta, \phi)$ is the total predicted sea-level change from these three ice sources. The symbols θ and ϕ denote co-latitude and east-longitude, respectively. We have confirmed, using a series of more complex ice balance models for the Antarctic and Greenland^{16–19}, that the large-scale geometry of the computed sea-level redistribution is relatively insensitive to our assumption of a uniform ice mass variation over these regions.

Figure 1 confirms that present-day variations in polar ice mass will be accompanied by dramatic departures of sea level from eustasy. For the sake of discussion, let us assume that a melting event in Greenland contributes 1 mm yr⁻¹ of eustatic sea-level rise. In this case, sea level will fall in the vicinity of Greenland, and it will rise by less than 0.2 mm yr⁻¹ in Newfoundland, Britain and Fennoscandia (Fig. 1b). There is a strong north–south gradient throughout the Atlantic, northern Pacific and Mediterranean, with a maximum sea-level rise of ~1.3 mm yr⁻¹ obtained in the northern Pacific and southern Atlantic. (The asymmetry in the latitudinal position of the sea-level maxima in Fig. 1b is due to rotational perturbations associated with Greenland mass flux.) The dominant

cause of the non-uniform redistribution is self-gravitation in the surface mass load. Water will pile up in the near field of an ice mass as a consequence of gravitational attraction; as the ice melts the ocean will relax and water will tend to flow from the near field to the far field. The same physical effect is evident in the remaining two frames of Fig. 1: the ice mass variation near the South Pole will be accompanied by a significant north–south gradient in the Antarctic Ocean, whereas the sea-level response to the melting of mountain glaciers is characterized by regionally localized minima. In all cases, an expectation of eustasy is only met in spatially limited geographic bands. Thus, secular sea-level trends inferred from tide gauge records may, depending on their location, sample a sea-level redistribution due to polar ice melting that significantly departs from the 'eustatic' measure of the ice balance.

We have computed sea-level trends at ~1,000 sites in the database supplied by the Permanent Service for Mean Sea Level (PSMSL)²⁰ for each of the three mass flux scenarios in Fig. 1 (see Supplementary Information). PSMSL sites mainly reside in the Northern Hemisphere. In the case of Antarctic mass flux, sea-level change at most of the PSMSL sites ranges up to 20% larger than the eustatic value; the exceptions are the southern portions of Africa, Australia and South America, where the sites will sample a sea-level change markedly smaller than the eustatic value. The site-specific departures from eustasy are even more dramatic for Greenland ice mass variations; in this case, less than 30% of the sites will record a sea-level change that is within 10% of the eustatic value. Sea-level change induced by the variations of Greenland ice, as measured at sites in Spitsbergen, Iceland and Fennoscandia, hovers close to zero but increases toward the eustatic value southward from the UK to North Africa and from the Canadian east coast to Florida. There is a further dramatic departure from the eustatic value at sites in regions such as Alaska, Yukon and British Columbia. These sites are also close to some of the largest melting events in Meier's database¹³ and thus exhibit extreme departures from the eustatic value that is associated with these sources of meltwater.

Estimates of global sea-level rise based on the analysis of tide gauge records require stringent site selection in order to avoid, for example, bias associated with interdecadal fluctuations and the contaminating effects of tectonic deformations⁵. In Fig. 2 we show predictions of sea-level change at 23 sites included in a recent tide gauge analysis⁸. In the case of Antarctic melting, these sites sample a sea-level change that ranges from 80% to 120% of the eustatic value. The departure from the eustatic value is large for European sites (numbers 1–8) in the case of Greenland melting. Indeed, the mean value for these sites is just 40% of the eustatic value.

The question arises as to whether geographic trends in observed tide gauge rates may be used to constrain meltwater contributions from the main ice reservoirs. A recent independent study²¹ based on standard sea-level theory¹² has analysed ~1,000 sites in the Revised Local Reference PSMSL data set²⁰ for this purpose. We proceed by analysing the more carefully selected subset of sites treated in Fig. 2. Figure 3a shows secular sea-level rates at the 23 sites sampled in Fig. 2. The mean rate obtained from these sites is 1.5 ± 0.1 mm yr⁻¹, the weighted (by $1/\sigma^2$) root-mean-square value of the residuals (w.r.m.s.) about the mean is 0.4 mm yr⁻¹, and the χ^2 residual value per-degree-of-freedom (henceforth χ^2) about the mean is 2.7. In Fig. 3b we show the same rates corrected for glacial isostatic adjustment (GIA) using the combination of ice and Earth model adopted in a number of earlier studies^{4–6,8}. This model is characterized by the ICE-3G deglaciation history²² and an Earth model with lower mantle viscosity of 2×10^{21} Pa s. The mean sea-level rate in this case is 1.8 ± 0.1 mm yr⁻¹, in agreement with ref. 8, and the w.r.m.s. and χ^2 about this mean are 0.4 mm yr⁻¹ and 3.4, respectively. Thus, the GIA correction previously applied to these tide gauge rates does not improve the geographic consistency of the residuals. Finally, in Fig. 3c we show the rates in Fig. 3a corrected using a GIA model proposed²³ on the basis of fits to tide gauge

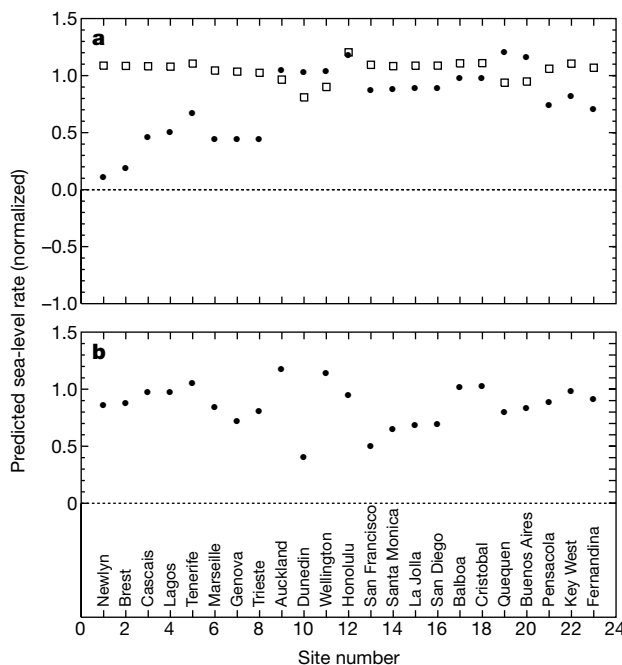


Figure 2 Predicted sea-level trends due to continuing ice mass variations at a subset of tide gauge sites. Shown are normalized sea-level rates from Fig. 1 for a selection of 23 sites (see labels at bottom) adopted in a recent tide gauge analysis⁸. (A final site, Lyttelton, New Zealand, is not included because it is characterized by an observed rate that is not consistent with rates at nearby sites). **a**, The square boxes and solid dots refer to predicted normalized sea-level rates for the case of Antarctic and Greenland ice mass variations. **b**, As in **a**, except for Meier's sources¹³.

records along the US east coast. This GIA model is identical to that used in Fig. 3b, with the exception that the lower mantle viscosity is increased to 5×10^{21} Pa s. In this case, the mean rate is 1.5 ± 0.1 mm yr⁻¹, and the w.r.m.s. and χ^2 about this mean are reduced to 0.3 mm yr⁻¹ and 2.1, respectively.

Next, we explore whether the geographic scatter evident in the GIA-corrected rates in Figs 3b and c can be reconciled on the basis of non-uniform sea-level change associated with polar melting events (Fig. 1). Following equation (1), we can denote the total sea-level change, at a site (θ_j, ϕ_j) , attributable to all sources by:

$$S_T(\theta_j, \phi_j) = V_O \times S_O^a(\theta_j, \phi_j) + S_I(\theta_j, \phi_j) \quad (2)$$

The first term on the right-hand side of equation (2) includes all contributions to sea-level change other (hence the subscript O) than the ice mass variations treated in Figs 1 and 2. Given the criteria used to select the 23 sites in Figs 2 and 3, we expect that these contributions will be dominated by ocean thermal expansion. Ideally, we can use equations (1) and (2) as the basis for a least-squares estimate of V_A , V_G and V_O , where the sea-level rates $S_T(\theta_j, \phi_j)$ are given by the 23 GIA-corrected tide-gauge rates in either Fig. 3b or c and the various normalized sea-level geometries $S^a(\theta_j, \phi_j)$ are known. The functions S_A^a , S_G^a and S_M^a are given in Fig. 1; however, the global geometry of ocean thermal expansion remains largely unknown.

As a preliminary application of the approach we proceed with a least-squares estimate of V_A , V_G and V_O by assuming that ocean thermal expansion contributes a eustatic sea-level rise (that is,

$S_O^a = 1$ over the oceans). Because the sea-level variation due to Antarctic mass flux is relatively flat for the 23 tide gauge sites we are considering (Fig. 2a), estimates for V_A and V_O will be highly correlated. Nevertheless, we have found that the least-squares procedure does yield a robust estimate of the sum $V_A + V_O$. The results of this calculation are shown by the solid lines superimposed on Fig. 3b and c. In Fig. 3b our best fit is obtained with $V_A + V_O = 0.99 \pm 0.13$ mm yr⁻¹ and $V_G = 0.54 \pm 0.15$ mm yr⁻¹. Thus the total eustatic value (including V_M) is ~ 2.0 mm yr⁻¹. The analogous expressions for Fig. 3c are $V_A + V_O = 0.61 \pm 0.13$ mm yr⁻¹ and $V_G = 0.60 \pm 0.15$ mm yr⁻¹ for a total eustatic value of ~ 1.7 mm yr⁻¹.

In Fig. 3c, the χ^2 of the observations about the solid line (1.1) is significantly better (at greater than 90% confidence) than the χ^2 relative to the simple mean of these rates (2.1). Even a qualitative inspection of Fig. 3c indicates that many of the residual trends evident in the GIA-corrected sea-level rates are remarkably well fitted by our best-fit (solid line) prediction for $S_T(\theta, \phi)$. Furthermore, the mean predicted rate for the European sites (numbers 1–8) is ~ 0.5 mm yr⁻¹ less than the mean predicted rate for the remaining sites. This difference is consistent with the observed trends on the same figure and provides a natural explanation for the long-standing observation that European sea-level rates are anomalously low^{8,24–26}.

The Meier¹³ tabulation of smaller glaciers is probably incomplete and subject to error. Furthermore, the mass balance of these glaciers may not be purely secular. We have repeated the least-squares

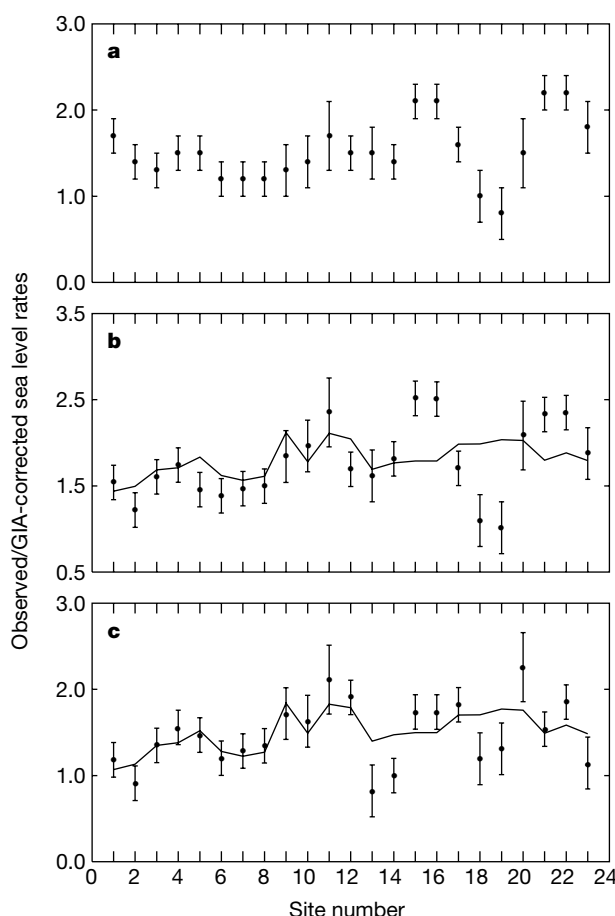


Figure 3 A new analysis of tide gauge records. **a**, Secular sea-level trends obtained (<http://www.pol.ac.uk/psmsl/>) from the PSMSL²⁰ tide gauge database for the set of 23 sites shown in Fig. 2. The error bars represent one- σ uncertainties. **b**, As in **a**, except for rates corrected using a GIA prediction based on an ice and Earth model combination adopted by previous studies^{4–6,8} (see text). **c**, As in **b**, except that we adopt a GIA correction advocated²³ on the basis of its fit to geographic trends evident in tide gauge

records from the US east coast. The solid line in frames **b** and **c** represents the best-fitting prediction through the GIA-corrected residual tide gauge rates determined by a least-squares fit of these residuals to equation (2). The parameters in this fit are the equivalent eustatic sea-level variation associated with present-day ice mass variations in the Antarctic (V_A) and Greenland (V_G) as well as 'other' sources (V_O) described in the text. In this exercise V_M is set to a value of 0.46 mm yr⁻¹, as prescribed in the Meier database¹³.

analysis in Fig. 3c by adopting a recent tabulation of 13 meltwater sources²⁷ in place of Meier's values for these sources (total V value of 0.34 mm yr^{-1}) and have found that our estimates are not significantly altered ($V_A + V_O = 0.74 \pm 0.13 \text{ mm yr}^{-1}$ and $V_G = 0.58 \pm 0.15 \text{ mm yr}^{-1}$). The former estimate is within the observational uncertainty cited in the second IPCC report¹ ($V_O = 0.2\text{--}0.7 \text{ mm yr}^{-1}$ and magnitude of $V_A < 1.4 \text{ mm yr}^{-1}$), while the latter is higher (by just over $1\text{--}\sigma$) than the upper bound cited for V_G (0.4 mm yr^{-1}).

Tide gauge records of sea-level change contain more information than has been recognized in previous studies of global sea-level rise. We have shown that geographic trends in (GIA-corrected) tide gauge rates at a set of carefully selected sites can be reconciled by invoking non-eustatic sea-level variations caused by continuing ice mass variations. We have furthermore demonstrated that it is possible to derive not only a single 'eustatic' measure of sea-level change but also, at least in principle, the independent contributions to this value from a subset of the various reservoirs of global ice. Our analysis is preliminary in several respects. First, ocean thermal expansion is clearly not a eustatic process¹, and our future work will incorporate a more realistic treatment of the process once its spatial geometry is better described²⁸. If significant latitudinal gradients in the geometry of thermal expansion exist within the Northern Hemisphere then these will influence our estimate of the amplitude of Greenland mass flux. Second, we have focused on a restricted set of tide gauge sites, and in future work we will cautiously extend our procedure in order to incorporate a larger number of robust tide gauge records. □

Received 14 September 2000; accepted 4 January 2001.

- Warrick, R. A., Le Provost, C., Meier, M. F., Oerlemans, J. & Woodworth, P. L. in *The Science of Climate Change: Contribution of Working Group I to the Second Assessment Report of the Intergovernmental Panel on Climate Change* (eds Houghton, J. T. et al.) 361–405 (Cambridge Univ. Press, Cambridge, 1990).
- Woodworth, P. L., Pugh, D. T., De Ronde, J. G., Warrick, R. A. & Hannah, J. (eds) *Sea Level Changes: Determination and Effects* (American Geophysical Union, Washington, 1992).
- Warrick, R. A., Barrow, E. M. & Wigley, T. M. L. (eds) *Climate and Sea Level Change: Observations, Projections and Implications* (Cambridge Univ. Press, Cambridge, 1993).
- Peltier, W. R. & Tushingham, A. M. Global sea-level rise and the greenhouse effect: Might they be connected? *Science* **244**, 806–810 (1989).
- Douglas, B. C. Global sea level rise. *J. Geophys. Res.* **96**, 6981–6992 (1991).
- Peltier, W. R. & Tushingham, A. M. Influence of glacial isostatic adjustment on tide gauge measurements of secular sea level change. *J. Geophys. Res.* **96**, 6779–6796 (1991).
- Trupin, A. S. & Wahr, J. M. Spectroscopic analysis of global tide gauge sea level data. *Geophys. J. Int.* **100**, 441–453 (1990).
- Douglas, B. C. Global sea level rise: A redetermination. *Surv. Geophys.* **18**, 279–292 (1997).
- Peltier, W. R. Postglacial variations in the level of the sea: Implications for climate dynamics and solid-Earth geophysics. *Rev. Geophys.* **36**, 603–689 (1998).
- Woodward, R. S. On the form and position of mean sea level. *US Geol. Surv. Bull.* **48**, 87–170 (1888).
- Daly, R. A. Pleistocene changes of sea level. *Am. J. Sci.* **10**, 281–313 (1925).
- Farrell, W. E. & Clark, J. T. On postglacial sea level. *Geophys. J. R. Astron. Soc.* **46**, 647–667 (1976).
- Meier, M. F. Contribution of small glaciers to global sea level. *Science* **226**, 1418–1421 (1984).
- Milne, G. A., Mitrovica, J. X. & Davis, J. L. Near-field hydro-isostasy: The implementation of a revised sea-level equation. *Geophys. J. Int.* **139**, 464–482 (1999).
- Dziewonski, A. M. & Anderson, D. L. Preliminary reference Earth model (PREM). *Phys. Earth Planet. Inter.* **25**, 297–356 (1981).
- James, T. S. & Ivins, E. R. Global geodetic signatures of the Antarctic ice sheet. *J. Geophys. Res.* **102**, 605–633 (1997).
- Bentley, C. R. & Giovinetto, M. B. in *Proceedings of the International Conference on the Role of the Polar Regions in Global Change* (eds Weller, G., Wilson, C. L. & Severin, B. A. B.) 481–488 (Univ. Alaska, Fairbanks, 1991).
- Jacobs, S. S., Hellmer, H. H., Doake, C. S. M., Jenkins, A. & Frolich, R. M. Melting of ice shelves and the mass balance of Antarctica. *J. Glaciol.* **38**, 375–387 (1992).
- Krabill, W. et al. Greenland ice sheet: High-elevation balance and peripheral thinning. *Science* **289**, 428–430 (2000).
- Spencer, N. E. & Woodworth, P. L. *Data Holdings of the Permanent Service for Mean Sea Level* (Permanent Service for Mean Sea Level, Bidston, Birkenhead, 1993).
- Plag, H.-P. & Jüttner, H.-U. Inversion of global tide gauges for present-day ice-load changes. *Proc. 2nd Int. Symp. Environ. Res. Arctic*. (in the press).
- Tushingham, A. M. & Peltier, W. R. Ice-3G: A new global model of late Pleistocene deglaciation based upon geophysical predictions of postglacial relative sea level. *J. Geophys. Res.* **96**, 4497–4523 (1991).
- Davis, J. L. & Mitrovica, J. X. Glacial isostatic adjustment and the anomalous tide gauge record of eastern North America. *Nature* **379**, 331–333 (1996).
- Shennan, I. & Woodworth, P. L. A comparison of late Holocene and twentieth-century sea level trends from the UK and North Sea region. *Geophys. J. Int.* **109**, 96–105 (1992).
- Woodworth, P. L., Tsimplis, M. N., Flather, R. A. & Shennan, I. A review of the trends observed in British Isles mean sea level data measured by tide gauges. *Geophys. J. Int.* **136**, 651–670 (1999).
- Lambeck, K., Smith, C. & Ekman, M. Tests of glacial rebound models for Fennoscandia based on

instrumented sea- and lake-level records. *Geophys. J. Int.* **135**, 375–387 (1998).

27. Trupin, A. S., Meier, M. F. & Wahr, J. M. Effect of melting glaciers on the Earth's rotation and gravitational field: 1965–1984. *Geophys. J. Int.* **108**, 1–15 (1992).

28. Levitus, S., Antonov, J. L., Boyer, T. P. & Stephen, C. Warming of the global ocean. *Science* **287**, 2225–2229 (2000).

29. Mitrovica, J. X. & Peltier, W. R. On postglacial geoid subsidence over the equatorial oceans. *J. Geophys. Res.* **96**, 20053–20071 (1991).

Acknowledgements

We thank J. Wahr, P. L. Woodworth and T. F. Baker for constructive reviews. We also thank H.-P. Plag for advice on the original manuscript and for sending us a preprint of his article with H.-U. Jüttner. M. Dyurgerov clarified recent models of mountain glacier mass balance. We acknowledge funding from the Ontario Government Premier's Research Excellence Award Program, the Canadian Institute for Advanced Research, NSERC, NASA, NSF, the Smithsonian Institution, NERC and the Royal Society of Great Britain.

Correspondence and requests for materials should be addressed to J.X.M. (e-mail: jxm@physics.utoronto.ca).

Geological constraints on tidal dissipation and dynamical ellipticity of the Earth over the past three million years

Lucas J. Lourens*, Rolf Wehausen† & Hans J. Brumsack†

* Faculty of Earth Sciences, Utrecht University, Budapestlaan 4, 3584 CD Utrecht, Netherlands

† Institute of Chemistry and Biology of the Marine Environment (ICBM), Carl von Ossietzky-University Oldenburg, PO Box 2503, D-26111 Oldenburg, Germany

The evolution of the Solar System has been shown to be chaotic¹, which limits our ability to retrace the orbital and precessional motion of the Earth over more than 35–50 Myr (ref. 2). Moreover, the precession, obliquity and insolation parameters^{3,4} can also be influenced by secular variations in the tidal dissipation and dynamical ellipticity of the Earth induced by glacial cyclicity^{3,5–10} and mantle convection¹¹. Here we determine the average values of these dissipative effects over the past three million years. We have computed the optimal fit between an exceptional palaeoclimate record from the eastern Mediterranean Sea and a model of the astronomical and insolation history³ by testing a number of values for the tidal dissipation and dynamical ellipticity parameters. We find that the combined effects of dynamical ellipticity and tidal dissipation were, on average, significantly lower over the past three million years, compared to their present-day values (determined from artificial satellite data and lunar ranging^{3,4,12}). This secular variation associated with the Plio-Pleistocene ice load history has caused an average acceleration in the Earth's rotation over the past 3 Myr, which needs to be considered in the construction of astronomical timescales and in research into the stationarity of phase relations in the ocean–climate system through time.

The main source of uncertainty in the computation of the precession and obliquity time series over the past three million years comes from changes in the Earth's dynamical ellipticity (E_D) due to surface mass load variations and tidal dissipation (TD) associated with the late Pliocene and Pleistocene ice-age cycles³. In particular, a change in E_D of about 0.00223 to that of the present-day could drive the precession and obliquity frequencies into resonance with a small perturbation term $s_6 - g_6 + g_5$ associated with Jupiter and Saturn³. Such a change in E_D can be reached during an ice age if the Earth would behave as a non-deformable (rigid) body³. But

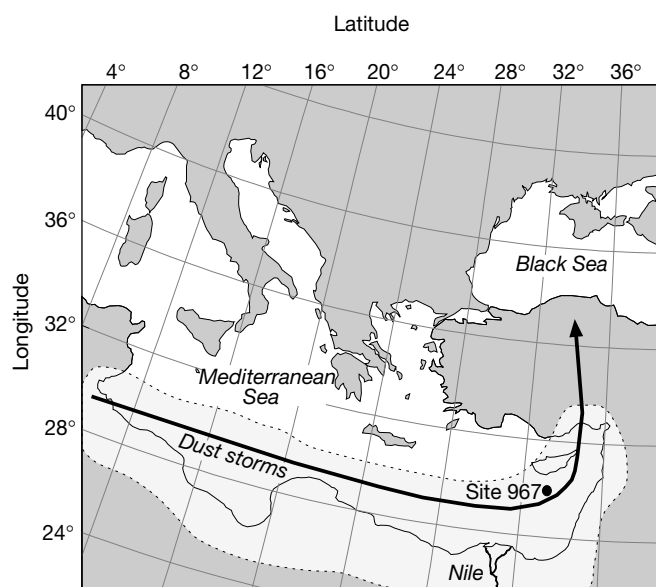


Figure 1 Location of ODP Site 967. The figure shows the trajectory of a typical dust storm observed in March 1984 (thick black line) and the area affected by dust²⁸ (area within thin broken line).

when the Earth's viscoelastic behaviour (for example, post glacial-rebound) is taken into account the proposed resonance seems unlikely^{7–10}. Nevertheless, ice ages will influence the astronomical parameters¹⁰, and ancient eclipse records for the past 2,700 years provide evidence for this¹². The eclipse study revealed an average value for the change in length of the mean solar day (ΔLOD) of ~ 1.7 ms per century, which is ~ 0.6 ms per century less than the present-day value. This increase has been attributed to a gradual decrease in the terrestrial oblateness owing to post-glacial (viscous) rebound processes^{9,12}.

Secular variations in the E_D or TD will alter the precession and obliquity periods, which have implications for past variations in insolation³. Hence, the extent of secular variations can be determined by examining palaeoclimate records that contain a mixture of precession and obliquity signals. Prime examples are characteristic sedimentary cycle (sapropels, carbonate cycles) patterns, related to precession-obliquity interference, in marine sequences of the Mediterranean Neogene¹³. A comparison between the patterns of sedimentary cycles and the insolation curves yielded a best fit for the astronomical solution of Laskar¹ (La90), with close to present-day values for E_D and TD¹³. These data are insufficient, however, to accurately determine a possible effect of ice ages on E_D or TD.

Recently¹⁴, an exceptional record of cyclic palaeoclimate variability was obtained from a section (75–86 m depth interval) of a core drilled at ODP site 967 in the eastern Mediterranean (Methods; Fig. 1). This interval contains a group of six sapropels (Fig. 2), which can be correlated cyclostratigraphically to time-equivalent sections elsewhere in the Mediterranean^{13,15,16}. We focused our investigations on the titanium to aluminium ratio as a proxy for climate change (Fig. 2). Both elements are important constituents of the terrigenous fraction. A significant part of the Al must have come from (suspended) Al-rich clay minerals, such as smectite¹⁷, supplied by the Nile River (Fig. 1). Titanium will be primarily of aeolian origin in (distal) marine sediments¹⁴. Dust storms occur in the rainy season between September and June, and are associated with strong upper winds of the subtropical jet stream (Fig. 1). We have interpreted the Ti/Al variability in terms of changes in the relative contribution of aeolian (for example, Sahara dust) versus fluvial (Nile) input (see also Methods). Low Ti/Al values indicate that relatively humid conditions prevailed at times of sapropel formation (that is, increased Nile discharge¹⁸), whereas high Ti/Al values point to more arid conditions in northern Africa during homogeneous marl deposition. The Ti/Al record shows cyclic variability throughout the studied interval, thus, also in the thick homogeneous intervals that separate large-scale sapropel clusters. The calibration of this record to summer insolation at 65°N is straightforward and

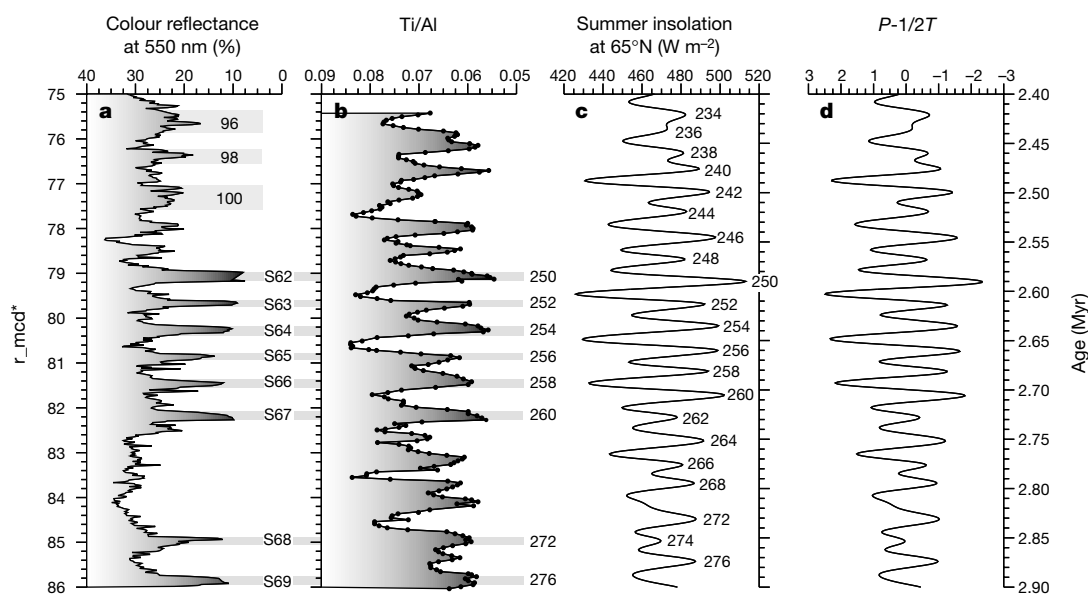


Figure 2 Cycle stratigraphy of the studied interval in ODP site 967. **a**, Colour reflectance values at 550 nm (%) against depth (see also Methods). Sapropels (S62–S69) are indicated by very low colour reflectance values. The expression of the glacial stages 100–96 (refs. 19) is also clearly visible between 75 and 78 m. **b**, Variations in Ti/Al against depth with on the right side the i-cycle codification¹³ for each sapropel. **c**, The 65°N

summer (average of May21 until July20) insolation curve (W m^{-2}) against age for solution La90_(1,1). Insolation maxima are labelled with their corresponding i-cycle codification¹³. **d**, The $P - 1/2T$ (normalized precession minus normalized tilt multiplied by 1/2) time series against age for solution La90_(1,1). Standard deviation and mean for precession and obliquity are calculated from the past 3 Myr.

unambiguous, and indicates that the studied time interval stretches from ~ 2.44 to ~ 2.90 Myr ago. This time interval characterizes a major amplification of dominant 41-kyr glacial variability, which is continuously present throughout the late Pliocene and early Pleistocene until it evolves into dominant 100-kyr rhythms during the late Pleistocene¹⁹.

Our procedure for computing average values for E_D and TD over the past 3 Myr is as follows. (1) We use a third-order polynomial fit to obtain midpoints of maxima and minima in the Ti/Al record. (2) We use solution La90 to calculate the precession and obliquity time series. This solution includes a full set of routines, which allow us to modify the E_D and TD terms³. We incorporate different values for γ ($= E_D/\nu$, where ν is the angular velocity of the Earth) ranging from 0.996 to 1.001 times that of the present-day value (J2000: $E_{D0} = 0.003280051$; $\nu_0 = 474659981.59757 \text{ arcsec yr}^{-1}$) with increments of 0.001. These different solutions were then calculated both with and without a correction for TD ($\dot{n}_{\text{mo}}/n_{\text{mo}} = -4.6 \times 10^{-18} \text{ s}^{-1}$, where n_{mo} is the sidereal motion of the Moon). Values adopted for the two parameters are indicated in subscript and as suffix to the astronomical solution. Thus, La90_(0.996,0) represents solution La90 with $\gamma = 0.996\gamma_0$ and no correction for the present-day tidal dissipation, while La90_(1,1) denotes solution La90 with present-day values for both parameters. In addition, we include the solution with $\gamma = \gamma_0$ and TD = 0.5 (that is, $\dot{n}_{\text{mo}}/n_{\text{mo}} = -2.3 \times 10^{-18} \text{ s}^{-1}$) as well as the outcome of two predictions of the Glacial Isostatic Adjustment Earth model with TD = 1. The first Earth model, La90_(RIGID,1) reflects the direct effect of surface mass load redistribution associated with the Plio-Pleistocene ice load history^{8,10}. The second prediction, La90_(FDW-GIA,1), is based on a viscoelastic Earth

model, which outcome fits very well with the constraints given by the eclipse records⁹. (3) We constructed $P-1/2T$ (normalized precession minus normalized tilt multiplied by 1/2) time series for each solution. The pattern in the $P-1/2T$ target curve is (almost) identical to that of the 65° N summer insolation curve (Fig. 2). We note that the choice for $P-1/2T$ instead of 65° N summer insolation does not influence the outcome of our study. (4) We determine ages of the minima and maxima for each solution with an accuracy of 100 years by fitting the original 1-kyr spaced astronomical series with a 100-kyr spaced sine curve. (5) Subsequently, we construct Ti/Al time series by assigning astronomical ages of $P-1/2T$ extremes to the correlative peaks in Ti/Al. This procedure implies small modifications of the Ti/Al time series for each solution (Fig. 3a, b). (6) We applied cross-spectral analyses (Fig. 3c) to calculate the precession and obliquity-related time lags between $P-1/2T$ and the Ti/Al time series for each solution (Fig. 4). (7) Finally, we calculate regression coefficients between the astronomical target curves and the (1-kyr interpolated) Ti/Al time series (Fig. 4).

The relationship between the obliquity-related time lags and the regression coefficients can be described by a third-order polynomial curve (Fig. 4). The optimum fit between Ti/Al and the $P-1/2T$ record is obtained for the La90_(1,0.5) solution (Fig. 4 main figure). But in fact, an infinity of solutions ranging between La90_(1.0003,0) and La90_(0.9997,1) will produce the same outcome. Nevertheless, our results indicate an average accelerative component in the Earth's rotation over the past 3 Myr, which is consistent with the trend derived from ancient records of solar and lunar eclipses over the past 2,700 years¹².

The present-day tidal dissipation factor, usually denoted by the

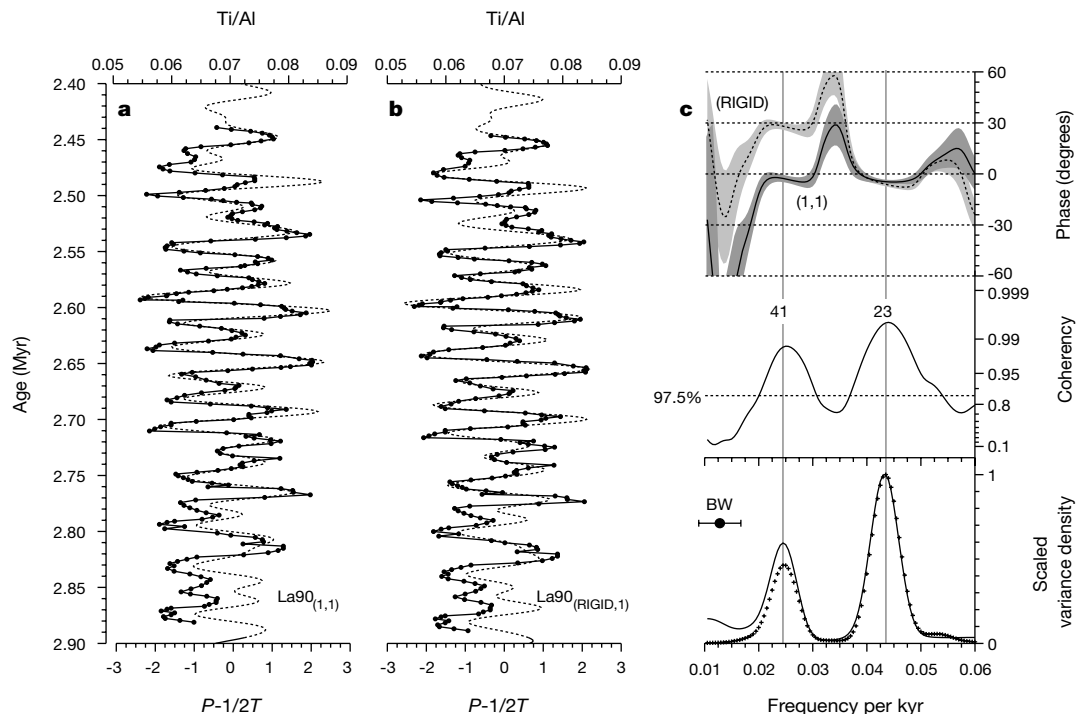


Figure 3 Examples of cross-spectra between Ti/Al and $P-1/2T$ time series.

a, Comparison between the time series for Ti/Al time (solid line) and $P-1/2T$ (dashed line) derived from the La90_(1,1) solution. **b**, Comparison between the time series for Ti/Al time (solid line) and $P-1/2T$ (dashed line) derived from the La90_(RIGID,1) solution. **c**, Lower panel: scaled variance density spectrum for Ti/Al time (solid line) and $P-1/2T$ (crosses) derived from solution La90_(1,1). Cross-spectral estimates³⁰ are based on 1-kyr linear interpolated time series for the time interval between 2.450 and 2.880 Myr ago. We used a Parzen smoothing-window with 238 lags, which resulted in a bandwidth of

0.0077910. Lower and upper confidence limits at the 97.5% level are given by: $0.385439 < \Delta P$ (97.5%)/ $P < 5.503866$. Middle panel: coherency spectrum plotted on an arc tan h scale. Top panel: phase spectrum (degrees) between the time series for Ti/Al and $p-1/2T$ derived from solutions La90_(1,1) and La90_(RIGID,1). We note that in the precession band phase lags between both solutions are (almost) the same, whereas an offset of ~ 30 degrees (~ 3.4 kyr) exists between both solutions in the obliquity band.

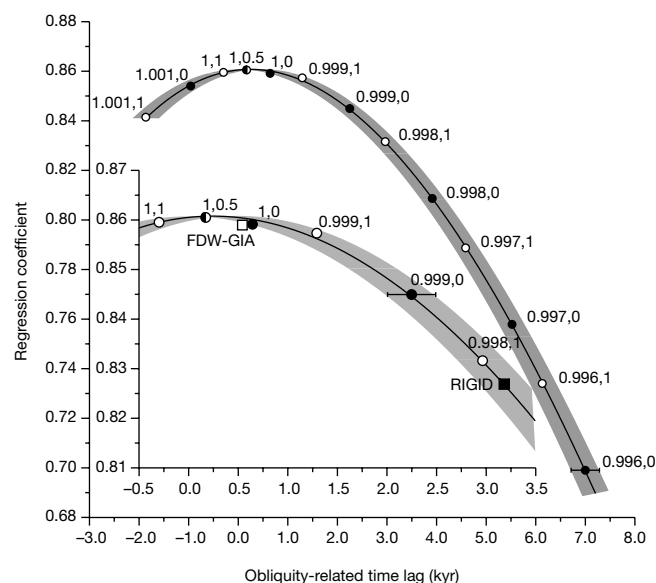


Figure 4 x - y plot of the regression coefficient against the obliquity-related time lag between the 1-kyr linear interpolated time series of Ti/Al and $P - 1/2T$ for all analysed La90 solutions (see text). **a**, Main figure, comparison for the La90 solutions, including different values for the dynamical ellipticity value (γ/γ_0) and tidal dissipation (TD). For example, 1.001, 1 stands for the La90 solution with $\gamma = 1.001\gamma_0$ and TD = 1. Filled circles indicate that TD = 0, open circles that TD = 1, and half-filled-half-open circle that TD = 0.5. The solid line is the outcome of a third-order polynomial fit. The shaded area indicates the error in the computed obliquity-related time lags associated with the 97.5% confidence limits. Inset, Magnified view of part of **a**. Included are the results obtained from two Earth model predictions¹⁰, La90_(FDW-GIA,1) (open square) and La90_(RIGID,1) (filled square).

ΔLOD or the acceleration of the Moon ($\dot{n}_{\text{mo}}$), has been derived from artificial-satellite data and lunar laser ranging estimates¹². In La90, a value for $\dot{n}_{\text{mo}}$ of -25.1 ± 1.3 arcsec per century² has been included, which corresponds with a ΔLOD of $+2.3 \pm 0.2$ ms per century (Methods). Since 1992, lunar laser ranging resulted in an estimate for $\dot{n}$ of -25.9 ± 0.5 arcsec per century², which approximates the -26.0 ± 2 arcsec per century² obtained from telescopic observations of transits of the planet Mercury across the solar disk and occultations of stars by the Moon from 1677 to 1973 (ref. 12). All this suggests that the lunar tidal acceleration was fairly constant over at least the past 300 years, although fluctuations in the Earth's rate of rotation at the millisecond level have been detected on a wide range of timescales from days to millennia¹².

According to our findings, the ΔLOD will be reduced to approximately $+1.2$ ms per century (compare $\dot{n} = -12.5$ arcsec per century²) on average over the past 3 Myr. Tidal friction takes place mainly in the hydrosphere, and in particular in shallow seas²⁰. Part of the reduced ΔLOD value can therefore be explained by a reduction in global tidal friction during periods of glacio-eustatic sea level lowering. Similarly, a part of the observed accelerative component of the Earth's rotation may be related to a glacial-induced change in E_D . The comparison with La90_(FDW-GIA,1) indicates, however, that this viscoelastic Earth model predicts a change in E_D that is too large with respect to our optimal fit (Fig. 4 inset). In particular, it is unlikely that the average tidal dissipation over the past 3 million years exceeded the present-day value, which is necessary to obtain an optimal fit between the La90 (FDW-GIA, 1) model and our data. Solution La90_(RIGID,1) produces a fit with our data that is less good, indicating that the effect of ice ages was too small to fulfill at least one of the conditions to perturb the Earth's precession and obliquity frequencies into the $s_6 - g_6 + g_5$ resonance^{3,8}.

We obtain an optimum fit between $P - 1/2T$ and the Ti/Al time series (Fig. 4) in the case where the obliquity-related time lag is approximately equal to that of precession ($\sim 150 \pm 300$ years), implying a constant and direct response of African aridity changes to insolation. Clearly, this assumption is critical for the outcome of the present study and contradicts others²¹. For instance, Clemens *et al.*²¹ assumed a 6-kyr lag for precession and a 13.4-kyr lag for obliquity (that is, a 7.4 kyr difference). According to this model, the best fit between our data and La90 will be obtained for $E_D < 0.996$ (with TD = 0), which on the basis of several viscoelastic Earth model predictions is highly unlikely¹⁰. Moreover, the time lags inferred are primarily based on late Pleistocene phase relations²². New highly accurate U/Th estimates indicate, however, that these time lags are overestimated considerably^{23,24}.

Secular variations in E_D and/or TD associated with the Plio-Pleistocene ice load history have caused an average accelerative component in the Earth's rotation over the past 3 Myr. This outcome needs to be considered in the construction of astronomical timescales and in research into the (non-) stationary nature of phase relations in the ocean-climate system through time. □

Methods

Ocean Drilling Program site 967 (34° 04' N, 32° 43' E) was drilled in the eastern Mediterranean near the Eratosthenes seamount at a water depth of 2,554 m. A composite depth profile (r_{mcd}) was constructed by correlating colour reflectance records (percentages at 550 nm wavelength) between different holes²⁵. The top 125 m of the site consists of early Pliocene to Holocene hemipelagic sediments, and contains 80 sapropels. These sapropels are not distributed evenly throughout the succession, but occur in small-scale and large-scale clusters. Four distinct large-scale clusters are found between 95 and 40 r_{mcd} . Biostratigraphic data¹⁵ indicate that these clusters correspond to the large-scale sapropel groups O, A, B and C as exposed in the land-based marine successions of the classical Vrica, Singa and Punta Piccola sections^{13,16}. These large-scale sapropel groups have been attributed to 400-kyr eccentricity maxima in the Earth's orbit¹⁶. The characteristic A group, which is found between 79 and 82.5 r_{mcd} in ODP 967, correlates to a 400-kyr eccentricity maximum in the Earth's orbit around 2.6 Myr ago¹⁶. Accordingly, the individual sapropels of this group were tuned to the 65° N summer insolation maxima, i-250 to i-260 (ref. 13).

Two holes, 967A and B, have been sampled. The continuity of the succession has been checked, using photographs and colour reflectance records of all holes. This resulted in a small modification of the reconstructed r_{mcd} values²⁵ for 967B-9H by adding 0.09 m (r_{mcd}^*). 159 samples were taken from 967B-9H-1, 0–1 cm to 9H-5, 28–29 cm, and 98 samples from 967A-8H-5, 0–1 cm to 8H-cc, 12–13 cm with a constant sample resolution of 4 cm. Before the analysis the samples were freeze-dried and then ground and homogenized in an agate ball mill. For X-ray fluorescence (XRF) analysis 600 mg of the sample powder were mixed with 3,600 mg lithium tetraborate ($\text{Li}_2\text{B}_4\text{O}_7$, Spektrumelt), pre-oxidized at 500 °C with NH_4NO_3 and fused to glass beads. The beads were analysed by a Philips PW 2400 X-ray spectrometer. Analytical precision was determined by parallel analysis of one international (GSR-6), and several in-house, standards and was better than 1% for Al and Ti.

The relative contributions of the fluvial and aeolian components to the Ti/Al record can be approximated, by applying some simple mass-balance calculations. The minimum value for Ti/Al, that is, 0.056 in sapropel 62 (Fig. 2), is close to the Ti/Al value of suspended matter in world rivers, which varies between 0.0485 and 0.0667 (refs 26, 27). The Ti/Al values decrease with increasing distance from the continent, due to preferential settling of heavy particles in the river mouth and on the continental shelf⁴. Most of the titanium found in the deep marine sediments of ODP Site 967 will therefore be of aeolian origin. At present, aeolian dust of African origin provides an important source of terrestrial input in the eastern Mediterranean²⁸. The dust trajectories parallel the North African coast and then turn north over Israel towards Turkey. The average Ti/Al value of particles from 23 dust storms in Israel between 1973 and 1988 arrives at 0.0876 (ref. 29). Assuming that all Nile-derived titanium settles in the delta (Ti/Al ≈ 0), this implies that $\sim 65\%$ of the terrigenous fraction is from aeolian origin during the formation of sapropel 62. The aeolian contribution may have increased to a maximum of $\sim 95\%$ during deposition of the homogeneous marls marked by the maximum Ti/Al value of ~ 0.084 (Fig. 2). A Ti/Al value of close to zero for the river-derived terrigenous component is a minimum value, and it is to be expected that this value is in fact larger (that is, between 0 and 0.056). Nevertheless, this will not alter the pattern and, as a consequence, variations in Ti/Al are interpreted as reflecting changes in the relative contribution of aeolian (for example, Sahara dust) and fluvial (Nile) sources.

Owing to the conserved angular momentum in the Earth-Moon system, the following empirical relation between $\dot{n}_{\text{mo}}$ and the retardation of the Earth's spin was found²⁰:

$$\dot{\Omega} = +(51 \pm 4)\dot{n}_{\text{mo}} \quad (1)$$

Following Stephenson and Morrison¹², the relationship between the acceleration of the Earth's spin expressed in units of rad s^{-2} and the ΔLOD in ms per century is:

$$1 \text{ ms per century} \equiv -2.67 \times 10^{-22} \text{ rad s}^{-2} \quad (2)$$

With $\dot{n}_{\text{mo}}$ being -25.1 ± 1.3 arcsec per century² in La90, and the relationships that $1 \text{ arcsec per century}^2 = 4.868 \times 10^{-25} \text{ rad s}^{-2}$, the present-day $\dot{\Omega}_{\text{tidal}}$ value is equivalent to a ΔLOD of $+2.3 \pm 0.2$ ms per century.

Received 6 April; accepted 11 December 2000.

- Laskar, J. The chaotic motion of the solar system: A numerical estimate of the size of the chaotic zones. *Icarus* **88**, 266–291 (1991).
- Laskar, J. The limits of Earth orbital calculations for geological time scale use. *Phil. Trans. R. Soc. Lond. A* **357** 1735–1759 (1999).
- Laskar, J., Joutel, F. & Boudin, F. Orbital precessional, and insolation quantities for the Earth from –20 to +10 Myr. *Astron. Astrophys.* **270**, 522–533 (1993).
- Quinn, T. R., Tremaine, S. & Duncan, M. A. A three million year integration of the Earth's orbit. *Astron. Journal* **101**, 2287–2305 (1991).
- Dehant, V., Loutre, M.-F. & Berger, A. Potential impact of the Northern Hemisphere Quaternary ice sheets on the frequencies of the astroclimatic orbital parameters. *J. Geophys. Res.* **95**, 7573–7578 (1990).
- Tushingham, A. M. & Peltier, W. R. ICE-3G: A new global model of late Pleistocene deglaciation based on geophysical predictions of post-glacial relative sea level change. *J. Geophys. Res.* **96**, 4497–4523 (1991).
- Peltier, W. R. & Jiang, X. The precession constant of the Earth: Variations through the ice age. *Geophys. Res. Lett.* **21**, 2299–2302 (1994).
- Mitrovica, J. X. & Forte, A. M. Pleistocene glaciation and the Earth's precession constant. *Geophys. J. Int.* **121**, 21–32 (1995).
- Mitrovica, J. X. & Forte, A. M. Radial profile of mantle viscosity: Results from the joint inversions of convection and postglacial rebound observables. *J. Geophys. Res.* **102**(B2), 2751–2769 (1997).
- Mitrovica, J. X., Forte, A. M. & Pan, R. Glaciation-induced variations in the Earth's precession frequency, obliquity and insolation over the last 2.6 Ma. *Geophys. J. Int.* **128**, 270–284 (1997).
- Forte, A. M. & Mitrovica, J. X. A resonance in the Earth's obliquity and precession over the past 20 Myr driven by mantle convection. *Nature* **390**, 676–679 (1997).
- Stephenson, F. R. & Morrison, L. V. Long-term fluctuations in the Earth's rotation: 700 BC to AD 1990. *Phil. Trans. R. Soc. Lond. A* **351**, 165–202 (1995).
- Lourens, L. J. *et al.* Evaluation of the Plio-Pleistocene astronomical timescale. *Paleoceanography* **11**, 391–413 (1996).
- Wehausen, R. & Brumsack, H.-J. Chemical cycles in Pliocene sapropel-bearing and sapropel-barren eastern Mediterranean sediments. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **158**, 325–352 (2000).
- Kroon, D. *et al.* Oxygen isotope and sapropel stratigraphy in the eastern Mediterranean during the last 3.2 million years. *Proc. ODP Sci. Res.* **160**, 181–189 (1998).
- Hilgen, F. J. Astronomical calibration of Gauss to Matuyama sapropels in the Mediterranean and implications for the geomagnetic polarity time scale. *Earth Planet. Sci. Lett.* **104**, 226–244 (1991).
- Venkatarathnam, K. & Ryan, B. F. Dispersal patterns of clay minerals in the eastern Mediterranean Sea. *Mar. Geol.* **11**, 261–282 (1971).
- Rossignol-Strick, M. African monsoons: an immediate climate response to orbital insolation. *Nature* **304**, 46–49 (1983).
- Raymo, M. E., Ruddiman, W. F., Backman, J., Clement, B. M. & Martinson, D. G. Late Pliocene variation in northern hemisphere ice sheets and North Atlantic Deep Water circulation. *Paleoceanography* **4**, 413–446 (1989).
- Lambeck, K. *The Earth's Variable Rotation: Geophysical Causes and Consequences* (Cambridge University Press, Cambridge, 1980).
- Clemens, S. C., Murray, D. W. & Prell, W. L. Nonstationary phase of the Plio-Pleistocene Asian monsoon. *Science* **274**, 943–948 (1996).
- Imbrie, J. *et al.* in *Milankovitch and Climate* (eds Berger, A., Imbrie, J., Hays, J., Kukla, G. & Saitzman, B.) 269–305 (Reidel, Dordrecht, 1984).
- Bard, E., Hamelin, B. & Fairbanks, R. G. U-Th ages obtained by mass spectrometry in corals from Barbados: sea level during the last 1300,000 years. *Nature* **345**, 405–410 (1990).
- Lawrence-Edwards, R., Chen, J. H., Ku, T.-L. & Wasserburg, G. J. Precise timing of the last interglacial period from mass-spectrometric determination of thorium-230 in corals. *Science* **236**, 1547–1553 (1987).
- Sakamoto, T., Janacek, T. & Emeis, K.-C. Continuous sedimentary sequences from the eastern Mediterranean sea: Composite depth sections. *Proc. ODP Sci. Res.* **160**, 37–59 (1998).
- Gordeyev, V. V. & Lisitsyn, A. P. The average chemical composition of suspensions in the world's rivers and the supply of sediments to the ocean by streams. *Dokl. Earth Sci.* **238**, 150–153 (1978). (Transl.)
- Martin, J.-M. & Meybeck, M. elemental mass-balance of material carried by major world rivers. *Mar. Chem.* **7**, 173–206 (1979).
- Ganor, E. & Mamane, Y. Transport of Saharan dust across the eastern Mediterranean. *Atmos. Environ.* **16**, 581–587 (1982).
- Ganor, E., Foner, H. A., Brenner, S., Neeman, E. & Lavi, N. The chemical composition of aerosols settling in Israel following dust storms. *Atmos. Environ.* **A 25**, 2665–2670 (1991).
- Paillard, D., Labeyrie, L. & Yiou, P. AnalySeries 1.1—Macintosh program performs time-series analysis. *Eos* **77**, 379 (1996).

Acknowledgements

We thank F. J. Hilgen, J. Laskar, J. X. Mitrovica, G. J. Reichert, J. Steenbrink and J. W. Zachariasse for discussions and comments, and J. Laskar and J. X. Mitrovica in particular for providing the astronomical solution and Earth model calculations, respectively. This work was supported by the Dutch National Science Foundation (NWO) and the German Science foundation (DFG) through ODP-SPP.

Correspondence and requests for materials should be addressed to L.J.L. (e-mail: llourens@geo.uu.nl).

Cranial design and function in a large theropod dinosaur

Emily J. Rayfield*, David B. Norman†‡, Celeste C. Horner‡, John R. Horner‡, Paula May Smith*, Jeffrey J. Thomason§ & Paul Upchurch*

* Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EQ, UK

‡ Museum of the Rockies, Montana State University, Bozeman, Montana, 59717, USA

§ Department of Biomedical Sciences, Ontario Veterinary College, University of Guelph, Guelph, Ontario, Canada, N1G 2W1

† Present address: Department of Paleobiology, National Museum of Natural History, Smithsonian Institution, Washington DC 20560, USA

Finite element analysis (FEA)¹ is used by industrial designers and biomechanicists to estimate the performance of engineered structures or human skeletal and soft tissues subjected to varying regimes of stress and strain^{2–4}. FEA is rarely applied to problems of biomechanical design in animals, despite its potential to inform structure–function analysis. Non-invasive techniques such as computed tomography scans can be used to generate accurate three-dimensional images of structures, such as skulls, which can form the basis of an accurate finite element model. Here we have applied this technique to the long skull of the large carnivorous theropod dinosaur *Allosaurus fragilis*⁵. We have generated the most geometrically complete and complex FEA model of the skull of any extinct or extant organism and used this to test its mechanical properties and examine, in a quantitative way, long-held hypotheses concerning overall shape and function^{6–8}. The combination of a weak muscle-driven bite force, a very ‘light’ and ‘open’ skull architecture and unusually high cranial strength, suggests a very specific feeding behaviour for this animal. These results demonstrate simply the inherent potential of FEA for testing mechanical behaviour in fossils in ways that, until now, have been impossible.

A 3D finite element model of the complete skull of the top Late Jurassic predator^{9–11} *Allosaurus fragilis* has been generated (Fig. 1a–c), using data from serial computed tomography (CT) scan images. This model has been loaded in order to simulate four different

Table 1 Bite force estimates for some living and extinct vertebrates

Species	Bite force	Calculation
<i>Allosaurus fragilis</i> biting mode A	805.42 N total	Low estimate bilateral static bite* at maxillary teeth 3, 4 and 5
<i>Allosaurus fragilis</i> biting mode B	2,147.88 N total	High estimate bilateral static bite at maxillary teeth 3, 4 and 5
<i>Allosaurus fragilis</i> biting mode C	18,746.76 N total	Maximum bilateral force at maxillary teeth 3, 4 and 5; with muscular and condylar force*
<i>Allosaurus fragilis</i> biting mode D	55,446.96 N total	Maximum bilateral force at maxillary teeth 3, 4 and 5; without muscular and condylar force*
<i>Allosaurus fragilis</i>	3,572.56 N	High estimate bilateral force at most posterior maxillary tooth 16
<i>Tyrannosaurus rex</i>	13,400 N†	Single tooth, possible unilateral bite, tooth impact velocity and adhering flesh accounted for
<i>Alligator mississippiensis</i>	13,000†	Unknown
<i>Panthera leo</i> (lion)	4,167.60 N‡	Calculated bilateral bite at molars
<i>Panthera pardus</i> (leopard)	2,268.7 N‡	Calculated bilateral bite at molars
<i>Felis concolor</i> (cougar)	1,836.8 N‡	Calculated bilateral bite at molars
<i>Canis lupus</i> (wolf)	1,412.2 N‡	Calculated bilateral bite at molars
<i>Vulpes vulpes</i> (red fox)	532.4 N‡	Calculated bilateral bite at molars

* Maximum tensile/compressive stress 200 MPa.

† Taken from ref. 12.

‡ Taken from ref. 28.

modes of biting (Loading conditions A to D; see Methods and Table 1). Biting modes A and B simulate a static bite generated only by the adductor (jaw closing) muscles (Fig. 1d). It is assumed that the skull is in equilibrium, with the jaws closing against a prey item. Dorsally directed bite forces were applied to the large maxillary teeth (nos 3–5, Fig. 1d) midway along the upper jaw. Corresponding forces were applied at the jaw joint (Fig. 1d) and at the origination sites of the principal jaw closing muscles (Fig. 1d, black arrowheads). Although not specifically a life situation, it was assumed that all adductor muscles were contracting maximally at the time of the bite to assess the maximum values of cranial stress.

For mode A, the applied bite force represents a minimum estimate based on an ectothermic model of dinosaur physiology (see Methods for calculation). Forces in mode B represent a

maximum estimate, using a homeothermic (endothermic) model (Table 1). The true bite force, condylar force and peak stresses experienced during a muscle-driven bite lie somewhere within the range of values calculated for modes A and B, exact values being a subjective matter based on personal views of dinosaur physiology.

In an attempt to estimate the strength of the skull, modes C and D represent the maximum forces that could be applied to the same impact teeth as in modes A and B before the cranium began to yield. To obtain these values, force magnitudes were increased iteratively until peak stresses reached yield point. In mode C the jaw adductors were considered to have been maximally active, and therefore tensioning the skull (Fig. 1d), whereas in mode D, the musculature was quiescent, so that the skull frame reacts passively to externally applied forces (parameters as in Fig. 1d, but without muscle and

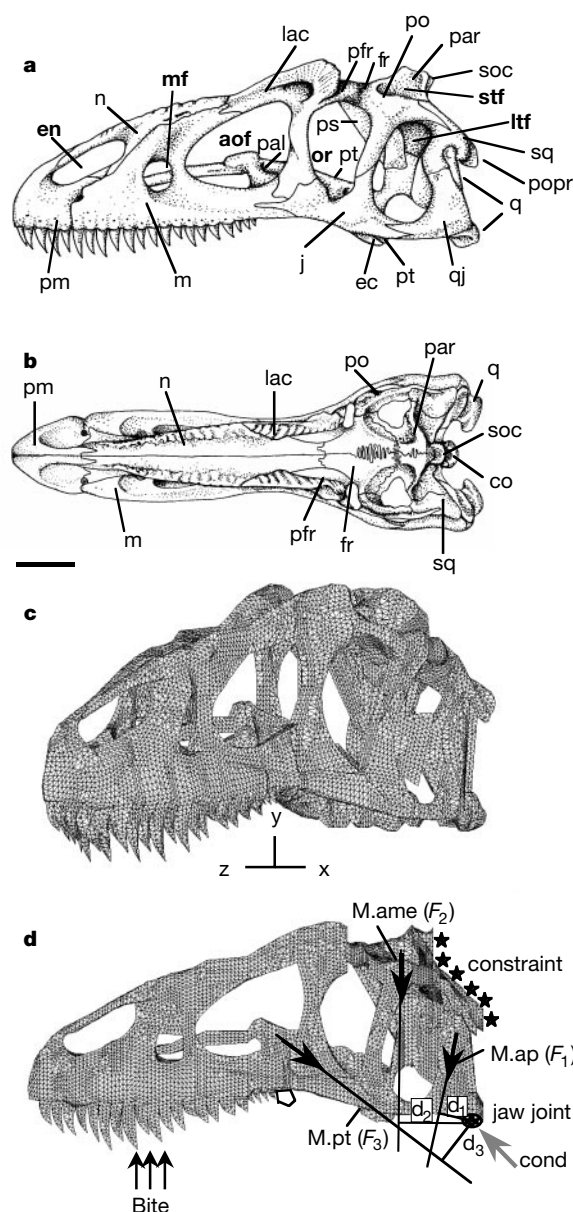


Figure 1 The skull of *Allosaurus fragilis*. **a**, Lateral view. **b**, Dorsal view. aof, antorbital fenestra; co, occipital condyle; ec, ectopterygoid; en, external naris; fr, frontal; j, jugal; lac, lacrimal; ltf, lower temporal fenestra; m, maxilla; mf, maxillary fenestra; n, nasal; or, orbit; pal, palatine; par, parietal; pfr, prefrontal; pm, premaxilla; po, postorbital; popr, paraoccipital process; ps, parasphenoid; pt, pterygoid; q, quadrate; qj, quadratojugal; soc, supraoccipital; sq, squamosal; stf, supratemporal fenestra. Scale bar, 10 cm.

c, Meshed finite element model of *Allosaurus fragilis* in oblique view. **d**, Meshed finite

element model in lateral view. Black stars, position of constraint; thick black arrows, line and direction of muscular force; thinner black arrows and 'bite', point of bite force; grey arrow, line and direction of condylar force; cond, point of condylar force; M.ap (F_1), M. adductor posterior; M.ame (F_2), M. adductor mandibulae externus superficialis, medius, profundus and M. pseudotemporalis; M.pt (F_3), M. pterygoideus group; d_1 , moment arm of muscle F_1 ; d_2 , moment arm of muscle F_2 ; d_3 , moment arm of muscle F_3 .

condylar force application). All models were anchored at the occipital surface on the rear of the skull (Fig. 1d, black stars), mimicking the stabilizing action of the vertebral column, cervical and axial musculature, and assuming that neck musculature was holding the head rigid rather than imparting any retractile forces.

Patterns of cranial stress distribution are similar for the four loading conditions. Stress magnitudes vary predictably, with higher magnitudes found where higher bite and jaw joint forces have been applied. The results of our analyses offer a number of hitherto unappreciated insights into the way the *Allosaurus* skull was designed functionally to resist bite- and impact-induced stresses in life.

First, compared with values for extant mammalian taxa, the muscle-driven bite force of *Allosaurus fragilis* appears to be relatively weak (Table 1). Also, *Allosaurus* adductor-generated bites are in no way comparable to estimated *Alligator mississippiensis* and *Tyrannosaurus rex* bite forces, both in the region of 13,000 N (ref. 12).

Second, in contrast to these weak muscular-driven bite forces, the skull is strong enough to withstand extremely large maximum forces at the tooth row before yielding because of tensional stress (Table 1, modes C and D; Fig. 2a, b). Although bone is much weaker in shear when tested mechanically, peak shear stresses in the skull are an order of magnitude lower than principal tension and compression. Should the dentition dislodge before skull failure, this would serve only to protect the skull from peak stresses making it more resilient. When all adductor musculature is contracting maximally and a condylar force is applied (mode C), the skull can withstand between 8.7 and 23.3 times the estimated muscle-driven bite force before yielding (Table 1, Fig. 2a, b). When the tensioning effect of muscular forces and the compressional effect of condylar forces are removed from the model (that is, mode D), the maximum force withstood by the tooth row may be up to 69 times the estimated muscle-driven bite force (Fig. 2b). Such strength of the skull means forces in the region of 55,000 N may be withstood by the skull (Fig. 2a, b). It appears that the skull is designed to be extremely strong when biting at the central maxillary teeth.

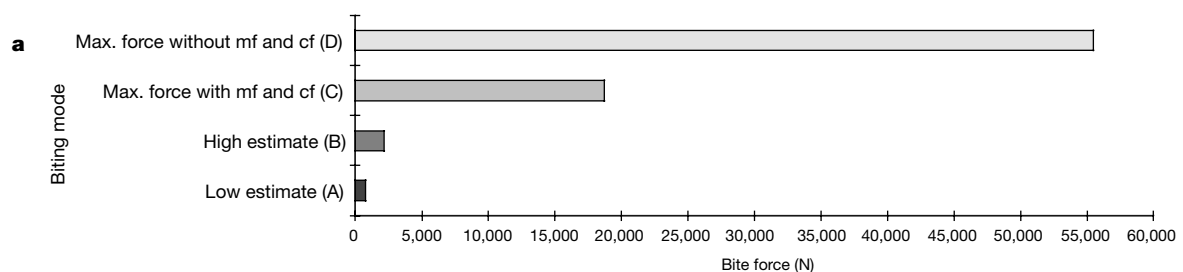
Third, tensile stresses ventrally and compressive ones dorsally reflect the bending moments acting on the skull during loading (Fig. 3a, b). However, the skull of *Allosaurus* seems to be designed to resist large vertically directed forces applied along the tooth row. In all biting modes, compressive stresses not absorbed by the robust

ventral maxilla (Fig. 1a, m) trace vector arcs from the impact teeth through the nasals and through the lateral struts forming the sidewall of the skull (Fig. 3b–d). Some vectors approach the skull roof, composed of thick, centrally located parietals and frontals (Fig. 1b, par, fr; Fig. 3d) that absorb a certain amount of stress. Many compressive vectors are routed in loops around the large cranial fenestrae (window-like openings in the skull) (Fig. 3b, c, d); these ‘functional loops’ minimize stress and strain in response to applied forces¹³. Such fenestrae (Fig. 1a, aof) appear to be important in stress management in the skull, and non-mechanical interpretations of the antorbital fenestra as a cavity housing a gland or air sac diverticulum¹⁴ appear to be of lesser importance.

Studies suggest that the large Cretaceous carnivore *T. rex*, with stout, conical teeth and an extremely robust skull¹⁵, could withstand tooth–bone impacts during feeding, and was capable of generating jaw-closure forces large enough to shatter skeletal material during prey dismemberment^{12,16,17}. A weak muscle-driven bite indicates that *Allosaurus* was not capable of splintering bony food material in this manner. In contrast to *T. rex*, *Allosaurus* displays recurved, laterally compressed teeth adapted for slashing or slicing⁸ and a more lightly constructed but extremely strong skull (Fig. 1a, b)⁵. We believe that this ‘weak bite/strong skull’ functional paradox may be explained by the predatory behaviour of *Allosaurus*.

Part of the overall high strength estimate for the allosaur skull can be accounted for by considering built-in safety factors; the ratio of a structure’s capacity or strength, compared with the highest expected load the structure experiences during everyday use¹⁸. Mammalian cranial bone may operate at a safety factor of between 1.8 and 11 (ref. 19). However, the very large difference between muscle-driven bite forces and the maximum force that can be applied to the skull before yielding (Fig. 2a, b) suggests that the *Allosaurus* cranium is overengineered. If the animal was to experience large forces such as those created by a high velocity impact of skull into prey, as part of its regular feeding strategy, the skull could still function within acceptable bounds of safety and the apparent overcompensation of design can be explained.

Our results provide quantitative evidence to suggest that during attack or feeding, *Allosaurus* generally used a high velocity impact of the skull into its prey; an analogue would be a person wielding a large, heavy hatchet. Aided by sharp, recurved teeth and powerful neck musculature driving the skull downwards and then imparting a retractile force, portions of flesh were sliced, torn away and



bite position	Range of adductor-generated bite force. mode A and B	Maximum bite force with muscle and condylar force applied. mode C	Ratio of adductor force: maximum force*	Maximum bite force without muscle and condylar force. mode D	Ratio of adductor force: maximum force†
Central	805 to 2,148 N	18,747 N	23.3 to 8.7	55,447 N	68.8 to 25.8

* Where maximum force = maximum bite force when muscular and condylar forces are applied

† Where maximum force = maximum bite force when muscular and condylar forces are not applied

Figure 2 Estimated bite forces and skull strength. **a**, Difference between muscular-generated bite force, modes A and B; and maximum force before yielding, modes C and D.

cf, condylar force; mf, muscle force. **b**, Table depicting ratios between muscular-generated bite force and maximum force in modes C and D respectively.

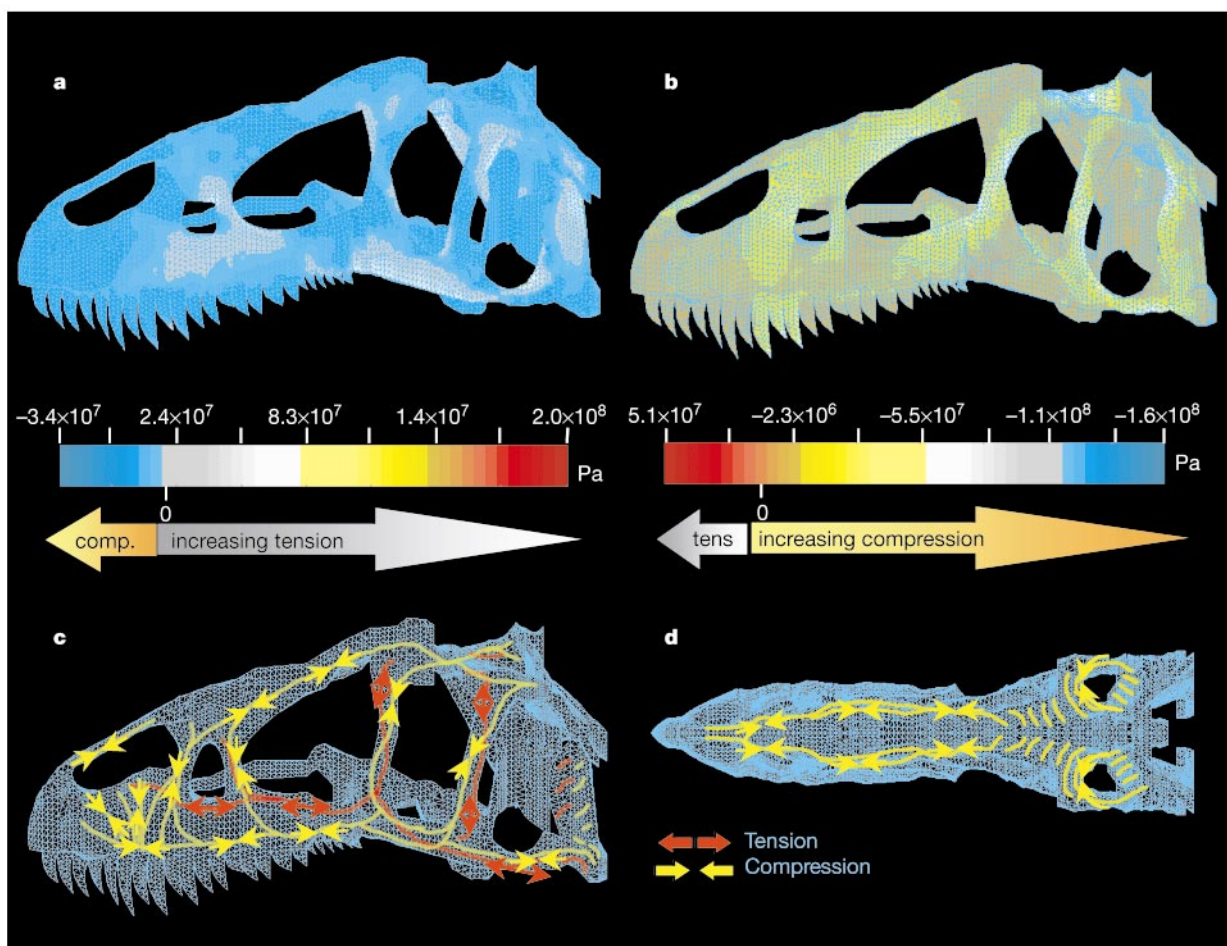


Figure 3 Stress distribution and vector plots for the skull of *Allosaurus fragilis* during a maximum impact bite without adductor muscle contraction (mode D). **a**, Stress distribution and magnitude plot for principal stress 1, lateral view. Colour scale bar indicates areas of high tension or compression. **b**, Stress distribution and magnitude plot

for principal stress 3, lateral view. See colour scale bar for areas of high tension or compression. **c**, Stylized plot showing stress vector direction, lateral view. **d**, Stylized stress vector plot, dorsal view.

swallowed (a strategy similar to that seen in *Varanus komodoensis*²⁰). The crushing bite of *T. rex* represents a specialization towards carcass dismemberment and possibly tackling larger, heavily armoured prey. By contrast, *Allosaurus* may have 'traded' a heavy skull and bite strength for greater speed and mobility of upper jaw impact in order to capture lighter and more agile forms such as ornithomimid dinosaurs²¹. *Allosaurus* might have ambushed larger, more dangerous, prey (for example, stegosaurs and sauropods) by inflicting a sudden devastating high-impact attack bite before the defender could retaliate.

The skull of *Allosaurus fragilis* is designed to allow this dinosaur to adopt a high impact 'slash and tear' mode of prey attack and feeding, rather than relying on a devastatingly powerful muscle-driven bite. The high cranial stresses associated with this mode of feeding are accommodated primarily by using a 'functional loop' model; this explains the unusually light and fenestrate architecture of the skull in an animal of this size. This analysis illustrates how, used with appropriate caution, FEA can be an important tool in analyses of mechanical behaviour in fossils. □

Methods

Skull mapping

An almost complete skull of *Allosaurus fragilis* from the Museum of the Rockies (MOR 693) was subjected to CT to obtain a series of transaxial scan images separated by 4-mm intervals. x,y coordinates from the CT images were imported into the finite element

modelling and analysis package COSMOS/M (version 2.0, Structural Research and Analysis Corp., CA, USA). Coordinates were used as a framework for the 3D geometry of *Allosaurus*. Resolution of CT scans was such that internal features, such as pneumatic cavities, could be identified and incorporated into the geometry of the model. The model was meshed creating solid four-noded tetrahedral elements (Fig. 1c). Number of elements = 146,398; nodes = 38,344; degrees of freedom = 107,421. Because the coordinates were calibrated initially, the model was to scale.

Material properties

Histologically, the bone of theropods most closely resembles the bone of fast-growing bovine mammals²². Studies of *Allosaurus* cranial bone²³ have shown it to be composed of secondary remodelled haversian bone with primary compact bone restricted to the surface²². Assuming that similar bone histology indicates broadly similar material properties, and to avoid overestimating maximum tension and compression, this model was assigned the material properties of bovine haversian bone: Young's modulus = 10 GPa; shear modulus = 3.6 GPa; poisson ratio = 0.4; density = 1.895 (ref. 23). The omission of compact bone from our model may lead to an underestimation of strength of the skull. However, inclusion of such strengthening characteristics would only serve to increase strength of the skull and provide further support for our conclusions concerning the difference between adductor-generated bite force and maximum force at failure. Properties of bovine dentine were applied to the teeth: Young's modulus = 21 GPa; shear modulus = 8 GPa; poisson ratio = 0.31; density = 2.076 (ref. 24).

Soft tissues

Adductor muscles were reconstructed in clay around a life-size cast of *Allosaurus*, and then incised at their widest part and the cross-sectional surface area measured. Cross-sectional areas were recorded digitally in Scion Image for Windows 95, 98 and NT, Beta 3b Version, an image analysis program for PC based on NIH image for Macintosh (<http://www.scioncorp.com/index.htm>). Adductor muscle force was calculated from muscle

stress values known from extant vertebrates. As these values range from 147 to 392 kPa (ref. 25), a low (mode A) and a high (mode B) estimate of possible muscle force were calculated using the extremes of this range. As stated in the text, the true muscle force values of *Allosaurus* lie within this range—exactly where depends on the animal's physiology. Ventrally directed muscle forces were applied at the attachment sites of all adductor muscles in the correct line of action based upon the anatomy of the lower jaw. Adductor muscles were grouped into three functional units. F_1 = M. adductor posterior; F_2 = M. adductor mandibulae externus group (comprising MAME superficialis, medialis and profundus) and M. pseudotemporalis; F_3 = M. pterygoideus anterior and posterior. The angles between lines of muscle action and the vertical were measured: for F_1 , $\alpha = 11^\circ$; for F_2 , $\beta = 3^\circ$; for F_3 , $\gamma = 62^\circ$.

Bite force calculation

To calculate a static, muscle-driven bite, it is assumed that all muscles are acting in a single parasagittal plane and that the skull is in equilibrium. In these models, *Allosaurus* is biting bilaterally at six teeth in total, the 3rd, 4th and 5th maxillary teeth, left and right sides (see Fig. 1d for details). Thus, three independent equations containing four unknowns are derived. One further assumption must be made; in this case that bite force is vertical. Equations calculate force on one side of the skull only, as forces are equal on both sides. The following equations are used (after refs 26, 27).

$$P \cos \theta + 3B = F_1 \cos \alpha + F_2 \cos \beta + F_3 \cos \gamma \quad (1)$$

$$P \sin \theta + F_1 \sin \alpha + F_2 \sin \beta = F_3 \sin \gamma \quad (2)$$

$$B[(\chi_1 + (\chi_1 + \chi_2) + (\chi_1 + \chi_2 + \chi_3))] = F_3 d_3 + F_2 d_2 + F_1 d_1 \quad (3)$$

Where P = condylar force, θ = angle of condylar force, $3B$ = total bite force at three adjacent teeth ($3B/3$ = bite force per tooth). F_1 , F_2 and F_3 = adductor muscle force values (low estimate: $F_1 = 228.88$ N; $F_2 = 1173.86$ N; $F_3 = 1,350.486$ N, high estimate: $F_1 = 610.3$ N; $F_2 = 3,130.51$ N; $F_3 = 3,601.32$ N). α , β , γ = angles from vertical for adductor muscle forces (as above). χ_1 = distance from jaw joint to max. 5; χ_2 = distance from max. 5 to max. 4; χ_3 = distance from max. 4 to max. 3; d_1 to d_3 = moment arms for muscle groups F_1 to F_3 , respectively; $d_1 = 0.0925$ m; $d_2 = 0.132$ m; $d_3 = 0.066$ m.

Using a high and a low estimate of muscle force leads to a high and a low estimate of bite force and condylar force (Table 1). Again, 'true' values lie within this range. By calculating such a range, assumptions concerning validity of loading parameters may be limited. Low estimate condylar force = 1,957.90 N per condyle; high estimate condylar force = 5,221.46 N per condyle. Angle of condylar forces from the vertical = 33.73° . Comparison with previously published bite force equations^{26,27} suggests that experimental error in the calculation of *Allosaurus* bite force is unlikely.

Finite element analysis calculates reaction to the applied load and a defined constraint for each element in turn, to give a composite picture of the mechanical behaviour of the skull (see Fig. 3).

For full descriptions of bite forces for all models featured in this analysis, see Table 1.

Received 24 March; accepted 22 November 2000.

1. Zienkiewicz, O. C. in *The Finite Element Method in Engineering Science* 521 (McGraw-Hill, London, 1971).
2. Ross, C. T. F. in *Finite Element Methods in Engineering Science* 519 (Ellis Horwood, London, 1990).
3. Huiskes, R. & Chao, E. Y. S. A survey of finite element analysis in orthopedic biomechanics: The first decade. *J. Biomechanics* **16**, 385–409 (1983).
4. Huiskes, R. & Hollister, S. J. From structure to process, From organ to cell: Recent developments of FE-analysis in orthopaedic biomechanics. *J. Biomech. Eng., Trans. ASME* **115**, 520–527 (1993).
5. Madsen, J. H. Jr *Allosaurus fragilis*: A Revised Osteology. *Utah Geol. Min. Surv. Bull.* **109**, 1–163 (1976).
6. Paul, G. S. in *Predatory Dinosaurs of The World* 464 (Simon and Schuster, New York, 1988).
7. Farlow, J. O. Speculations about the diet and foraging behaviour of large carnivorous dinosaurs. *Am. Mid. Nat.* **95**, 186–191 (1976a).
8. Farlow, J. O., Brinkman, D. L., Abler, W. L. & Currie, P. J. Size, shape, and serration density of theropod dinosaur lateral teeth. *Mod. Geol.* **16**, 161–198 (1991).
9. Dodson, P., Behrensmeyer, A. K., Bakker, R. T. & McIntosh, J. S. Taphonomy and paleoecology of the dinosaur beds of the Jurassic Morrison Formation. *Paleobiol.* **6**, 208–232 (1980).
10. Molnar, R. E. & Farlow, J. O. in *The Dinosauria* (eds Weishampel, D. B., Dodson, P. & Osmolska, H.) 210–224 (Univ. California Press, Berkeley, 1990).
11. Perez-Moreno, B. P. et al. On the presence of *Allosaurus fragilis* (Theropoda: Carnosauria) in the Upper Jurassic of Portugal: first evidence of an intercontinental dinosaur species. *J. Geol. Soc. Lond.* **156**, 449–452 (1999).
12. Erickson, G. M. et al. Bite-force estimation for *Tyrannosaurus rex* from tooth-marked bones. *Nature* **382**, 706–708 (1996).
13. Buckland-Wright, J. C. Bone structure and the patterns of force transmission in the cat skull (*Felis catus*). *J. Morph.* **155**, 35–62 (1978).
14. Witmer, L. M. The Evolution of the Antorbital Cavity of Archosaurs: A study in soft-tissue reconstruction in the fossil record with an analysis of the function of pneumaticity. *J. Vert. Paleo. Mem.* **17**, 1–73 (1997).
15. Molnar, R. E. The cranial morphology of *Tyrannosaurus rex*. *Palaeontographica Abr. A.* **217**, 137–176 (1991).
16. Chin, K., Tokaryk, T. T., Erickson, G. M. & Calk, L. C. A king-sized theropod coprolite. *Nature* **393**, 680–682 (1998).

17. Erickson, G. M. & Olson, K. H. Bite marks attributable to *Tyrannosaurus rex*: preliminary description and implications. *J. Vert. Paleontol.* **16**, 175–178 (1996).
18. Diamond, J. M. in *Principles of Animal Design: The Optimization and Symmorphosis Debate* (eds Weibel, E. R., Taylor, C. R. & Bolis, L.) 21–27 (Cambridge Univ. Press, Cambridge, 1998).
19. Thomason, J. J. & Russell, A. P. Mechanical factors in the evolution of the mammalian secondary palate: a theoretical analysis. *J. Morph.* **189**, 199–213 (1986).
20. Auffenberg, W. *The Behavioural Ecology of the Komodo Monitor*. (Univ. Florida Press, Gainesville, Florida, 1981).
21. Sues, H. D. & Norman, D. B. in *The Dinosauria* (eds Weishampel, D. B., Dodson, P. & Osmolska, H.) 498–509 (Univ. California Press, Berkeley, 1990).
22. Reid, R. E. H. Bone histology of the Cleveland-Lloyd dinosaurs and of dinosaurs in general. *Brig. Young Univ. Geol. Stud.* **41**, 25–71 (1996).
23. Reilly, D. T. & Burstein, A. H. The elastic and ultimate properties of compact bone tissue. *J. Biomech.* **8**, 393–405 (1975).
24. Waters, N. E. in *The Mechanical Properties of Biological Materials* (eds Vincent, J. F. V. & Currey, J. D.) 99–135 (Cambridge Univ. Press, Cambridge, 1980).
25. Thomason, J. J., Russell, A. P. & Morgelli, M. Forces of biting, body size, and masticatory muscle tension in the opossum *Didelphis virginiana*. *Can. J. Zool.* **68**, 318–324 (1990).
26. Crompton, A. W. & Hylander, W. L. in *The Ecology and Biology of Mammal Like Reptiles* (eds Hotton, N., Maclean, P. D., Roth, J. J. & Roth, E. C.) 263–282 (Smithsonian Institution Press, Washington DC, 1986).
27. Sinclair, A. G. & Alexander, R. M. Estimated forces exerted by the jaw muscles of some reptiles. *J. Zool.* **213**, 107–115 (1987).
28. Thomason, J. J. Cranial strength in relation to estimated biting forces in some mammals. *Can. J. Zool.* **69**, 2326–2333 (1991).

Acknowledgements

We are grateful to the Bozeman Hospital, Montana, USA, and Picker CT scan Company for CT scanning and file conversion; R. E. H. Reid for advice on dinosaur bone histology and material properties; and M. Harwood for COSMOS/M technical support. J. R. Horner provided access to *Allosaurus fragilis* (MOR 693), C. C. Horner provided use of computing equipment and aided file conversion; P. May Smith assisted with bite force calculations; S. Evans provided crocodile and alligator material for dissection and R. Felix advised on coordinate capture and image production. This work was supported by The Royal Society (D.B.N.), the Natural Environment Research Council, The Department of Earth Sciences and Emmanuel College, Cambridge University and the Cambridge Philosophical Society, and the British Federation of Women Graduates (E.J.R.).

Correspondence and requests for material should be addressed to E.J.R. (email: eray@esc.cam.ac.uk).

Genetic evidence against panmixia in the European eel

Thierry Wirth & Louis Bernatchez

GIROQ, Département de Biologie, Université Laval, Ste-Foy, Québec G1K7P4, Canada

The panmixia hypothesis—that all European eel (*Anguilla anguilla*) migrate to the Sargasso Sea for reproduction and comprise a single, randomly mating population—is widely accepted^{1,2}. If true, then this peculiar life history strategy would directly impact the population genetics of this species, and eels from European and north African rivers should belong to the same breeding population through the random dispersal of larvae. To date, the panmixia hypothesis has remained unchallenged: genetic studies realized on eel's mitochondrial DNA failed to detect any genetic structure^{3–5}, and a similar lack of structure was found using allozymes^{6,7}, with the exception of clinal variation imposed by selection^{8,9}. Here we have used highly polymorphic genetic markers that provide better resolution^{10,11} to investigate genetic structure in European eel. Analysis of seven microsatellite loci among 13 samples from the north Atlantic, the Baltic Sea and the Mediterranean Sea basins reveals that there is global genetic differentiation¹². Moreover, pairwise Cavalli-Sforza and Edwards¹³ chord distances correlate significantly with coastal geographical distance. This pattern of genetic structure implies

non-random mating and restricted gene flow among eels from different sampled locations, which therefore refute the hypothesis of panmixia. Consequently, the reproductive biology of European eel must be reconsidered^{14,15}.

Glass eels or fins from adults were collected from different rivers across the European and African coasts (Fig. 1) during spring and autumn 1999. Whole-cell DNA was extracted from 611 individuals and subsequently genotyped at seven mendelian-inherited microsatellite loci. All loci were highly polymorphic in all sampled localities, and the average number of alleles per locus per sample ranged from 12.92 (± 1.44) to 21.23 (± 2.59). Observed and expected mean heterozygosities per sample ranged from 0.835 (± 0.064) and 0.874 (± 0.016) to 0.895 (± 0.054) and 0.934 (± 0.008), respectively. Probability tests of Hardy–Weinberg equilibrium (HWE) using a Markov chain approach (10,000 iterations) were conducted for each of the 13 samples¹⁶. Seven out of ninety-one cases showed significant departure from HWE after Bonferroni¹⁷ correction ($\alpha = 0.05$, $k = 13$). These departures were not clustered by locus or sample, although the locus Aro121 was involved in four significant tests.

The genetic differentiation based on allelic frequency distribution over all samples was highly significant, and revealed unexpected relationships among the so-called ‘panmictic eel’ samples ($P = 0.0017$; 10,000 iterations). Population structure among these samples was confirmed by a weak but significant global genetic differentiation (F_{ST}) value of 0.0017 ($P = 0.0014$; 10,000 iterations). Moreover, a pairwise F_{ST} matrix resulted in significant values for south (Mediterranean Sea samples) versus north (Baltic and North Sea) comparisons (6 cases; range 0.003–0.005) and for the Motala river versus other sample comparisons (8 cases; range 0.005–0.012). In the latter comparisons, the smaller sample size for the Motala river ($n = 24$) may be responsible for the significant tests of differentiation.

A principal challenge in situations of slight genetic differentiation is to discriminate between minor but real population structure^{18–20} and artefacts due to noise related to sampling errors. Further evidence for population structuring was obtained here by plotting pairwise D_{CE} values against coastal distances (Fig. 2). The Pearson's correlation (r) between the two factors was highly significant ($P < 0.0045$; Mantel test²¹), and the linear regression model explained 21% of the variance between them. This result is therefore con-

sistent with the hypothesis that the European eel exhibits isolation by distance, which implies non-random mating and restricted gene flow among eels from different sampled locations. We also analysed the effect of a single locus on this correlation (r) and the significance (P) of the regression of D_{CE} versus coastal distance by jack-knifing (Table 1). The r values were slightly smaller and the P values slightly less significant than estimated with all seven loci; however, the similar patterns shown by each jack-knifed data set imply that no single locus determined the overall observed positive correlation between D_{CE} and coastal distance. Consequently, selection mediated through gene linkage for particular microsatellite loci could be dismissed. To test whether a bias had been introduced by the inclusion of the Icelandic sample (geographically distant population, putative hybridization with the American eel *A. rostrata*), we redid the analysis without this sample. The results proved to be robust as the correlation between population differentiation and distance remained highly significant ($r = 0.462$ and $P < 0.007$).

To illustrate the putative geographical pattern of genetic differentiation a phenogram was constructed from the D_{CE} pairwise distance matrix using a neighbour-joining²² algorithm (Fig. 3). We added samples of American eel for the purpose of rooting the tree (our unpublished data). The phenogram grouped the Mediterranean samples in a distinct clade, as well as the North Sea and Baltic sea samples, despite weak bootstrap values. Icelandic eels were intermediate between other *A. anguilla* samples and *Anguilla rostrata*, which is congruent with the genetic status of Icelandic eels for which hybridization with the American eel has been documented²³.

Table 1 Pearson's correlation (r) and the significance (P) of the regression of D_{CE} versus coastal geographical distance in km

Locus excluded*	r	P
Aro 054	0.467	< 0.007
Aro 121	0.298	< 0.028
Ang 114	0.403	< 0.009
Aro 095	0.431	< 0.012
Aro 063	0.533	< 0.014
Ang 151	0.388	< 0.012
Ang 101	0.445	< 0.008
Mean	0.424	
Standard error	0.027	

Based on data sets from which one locus had been excluded by a jack-knifing procedure.



Figure 1 Sampling locations. Locations where glass eels muscle was sampled: Adour, river ($n = 50$); Couesnon, river ($n = 50$); Grand-Lieu, lake ($n = 50$); Minho, river ($n = 50$); Moulonya, oued ($n = 50$); Salses-Leucate, laguna ($n = 41$); Severn, estuary ($n = 50$); Tiber, river ($n = 50$). Locations where yellow eels fin clips were sampled: Arreso, lake ($n = 46$); Elbe, river ($n = 50$); Imsa, river ($n = 50$); Motala, river ($n = 24$); Olfus, river ($n = 50$).

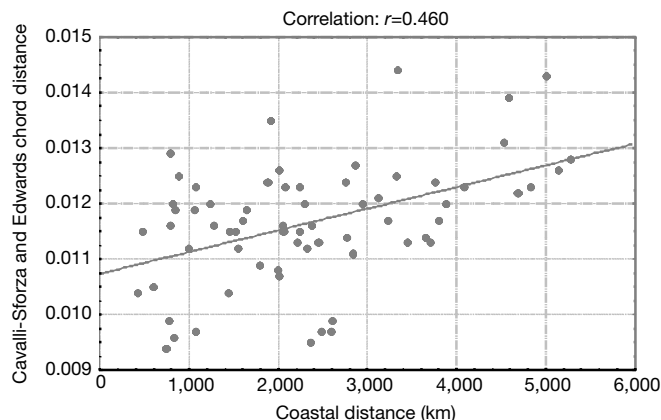


Figure 2 Genetic distance (D_{CE}), based on seven microsatellite loci, versus coastal geographical distances for all possible pairwise combinations of 13 *A. anguilla* samples. A significant positive correlation is observed. Regression analysis: $y = 0.0107 + 4 \times 10^{-7}x$, $r^2 = 0.212$, $P = 0.0045$; Mantel test. A significant positive correlation indicates that samples are spatially genetically structured, with isolation by distance playing an important role. To avoid sample size effects, the Motala sample was not included in this analysis.

Our results thus clearly show that a weak but highly significant genetic structure exists in the European eel, which seems to be related, at least partially, to the maritime units. The panmixia paradigm for this species must therefore be reconsidered. We now have to determine which one of the alternative life history models best explains the observed genetic structure. In a first model, there is a temporal delay between the arrival of adult eels from different latitudes at the common breeding site, which induces higher similarities of synchronic samples breeding together and subsequently larger genetic distances when compared with diachronic samples. North Atlantic individuals may start their reproductive phase earlier in the spring, as the distance to migrate back to the Sargasso Sea is the shortest. This would be followed by North Sea and Baltic Sea individuals and then Mediterranean eels. This temporal allopatry should be reinforced by non-random return of larvae to their parents' ocean basin through active swimming²⁴ of the leptocephali or through seasonal current changes²⁵.

In a second model, more than one reproductive area is used by different populations and different currents carry the leptocephali back to their parent's original freshwater habitat. Third, there is only one shared spawning area where assortative mating occurs and larval homing to parents' habitat takes place using an unknown mechanism. Alternatively, homing can take place at later developmental stages involving active juvenile saltwater dispersal.

We concur that the first hypothesis is the most likely, and this does not cast doubt on the observations of Schmidt¹. The duration of the spawning season (late winter to early summer¹) might therefore correspond to the arrival of successive waves of reproductive migration. Alternatively, the acceptance of the second hypothesis implies that field observations of Schmidt failed to locate a second or third reproductive area, which seems to be unlikely. Finally, the last hypothesis invokes assortative mating, but the relative important rates of hybridization between American and European eels²³ (5% in Iceland) weakens the efficiency of such processes at the intraspecific level.

More detailed and local ecological and genetic investigations will now be necessary in order to elucidate the exact ecological processes involved in shaping the pattern of population structuring we reported in this study. □

Methods

Genotyping

Genomic DNA was isolated from glass eels muscle or from adults' fin clips according to standard methods²⁶. The microsatellite flanking sequences and primers are available on

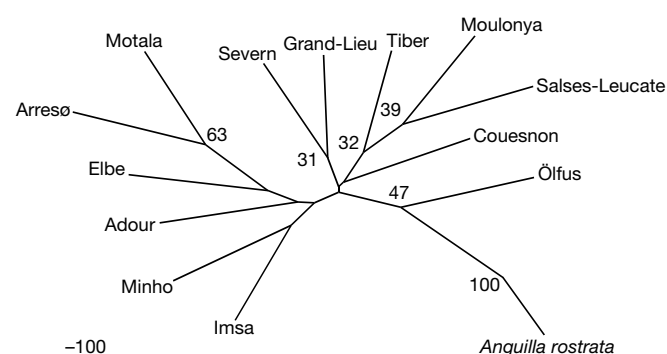


Figure 3 Neighbour-joining phenogram depicting genetic distance relationships based on Cavalli-Sforza and Edwards' chord distances among 13 European eel samples. American eel was used as an outgroup. Values on the nodes represent the percentage of bootstrap replicates over loci ($n = 2,000$) showing the particular nodes, only those with values higher than 30% are reported. Branch lengths are proportional to the genetic distance between the taxa. The scale bar represents a Cavalli-Sforza and Edwards distance (D_{CE}) of 0.0003.

GenBank under the accession numbers AF237896–AF237902. Polymerase chain reactions (PCRs) were performed in duplex or triplex with polymerase and rhodamine-marked primers (Perkin Elmer). Reactions were pooled and alleles belonging to seven different loci were segregated on an ABI377 automated sequencer. The sizes of the fragments were determined in reference to a size standard running in each lane using the software Genescan version 2.1 and Genotyper version 2.0.

Genetic variability and population genetics parameters

Allelic diversity, genetic variation (observed heterozygosity and unbiased heterozygosity under HWE), deviation from HWE and genetic differentiation were calculated with GENEPOP. Variation of allelic frequencies among samples of *A. anguilla* was assessed by first testing the null hypothesis of homogeneity in allelic distribution by Fisher's exact test using the Markov chain method and then by quantifying the standardized variance in allelic frequencies (θ) as an estimator of F_{ST} , using a described method¹² as implemented in GENETIX 4.0.

Isolation by distance

The relationship of genetic divergence to geographical separation of sites was examined by measuring the Cavalli-Sforza chord distance¹³ on the basis of allelic frequencies across the seven loci between each pair of samples. The significance of the relationships between D_{CE} and geographical distances cannot be evaluated using standard regression techniques, as the regression is based on non-independent, pairwise comparisons. We used Mantel's test²¹ to assess the significance of the observed correlations. To determine whether one locus contributed disproportionately to the geographic pattern of genetic differentiation, we excluded loci one at a time (jack-knifing over loci) and recalculated the Pearson's correlation (r) of D_{CE} versus geographical distance.

Phylogenetic inference

Cavalli-Sforza and Edwards' chord distance was used to construct a phylogenetic tree using a neighbour-joining algorithm²⁷. Support for the tree nodes was assessed by bootstrapping over loci (5,000 iterations). The tree was built using PHYLIP²⁷ version 3.5c from raw allelic frequencies.

Received 5 June; accepted 4 December 2000.

- Schmidt, J. The breeding places of the eel. *Smithsonian Inst. Ann. Rep.* **1924**, 279–316 (1925).
- Tesch, F.-W. *The Eel* (Chapman and Hall, London, 1977).
- Avise, J. C., Helfman, G. S., Saunders, N. C. & Hales, L. S. Mitochondria DNA differentiation in North Atlantic eels: population genetic consequences of an unusual life history pattern. *Proc. Natl Acad. Sci. USA* **83**, 4350–4354 (1986).
- Sang, T.-Z., Chang, H.-Y., Chen, C.-T. & Hui, C.-F. Population structure of the Japanese eel, *Anguilla japonica*. *Mol. Biol. Evol.* **11**, 250–260 (1994).
- Lintas, C., Hirano, J. & Archer, S. Genetic variation in the European eel (*Anguilla anguilla*). *Mol. Mar. Biol. Biotech.* **7**, 263–269 (1998).
- DeLigny, W. & Pantelouris, E. M. Origin of the European eel. *Nature* **246**, 518–519 (1973).
- Comparini, A., Rizzotti, M. & Rodinó, E. Genetic control and variability of phosphoglucose isomerase (PGI) in eels from the Atlantic ocean and the Mediterranean sea. *Mar. Biol.* **43**, 109–116 (1977).
- Pantelouris, E. M., Arnason, A. & Tesch, F. W. Genetic variation in the eel. *Genet. Res. Camb.* **16**, 277–284 (1970).
- Williams, G. C. & Koehn, R. K. in *Evolutionary Genetics of Fishes* (ed. Turner, B. J.) 529–560 (Plenum, New York, 1984).
- Ruzzante, D. E., Taggart, C. T., Cook, D. A. & Goddard, S. V. Genetic differentiation between inshore and offshore Atlantic cod (*Gadus morhua*) of Newfoundland: a test and evidence of temporal stability. *Can. J. Fish. Aquat. Sci.* **54**, 2700–2708 (1998).
- Shaw, P. W., Turan, C., Wright, J. M., O'Connell, M. & Carvalho, G. R. Microsatellite DNA analysis of population structure in Atlantic herring (*Clupea harengus*), with direct comparison to allozyme and mtDNA RFLP analyses. *Heredity* **83**, 490–499 (1999).
- Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).
- Cavalli-Sforza, L. L. & Edwards, A. W. F. Phylogenetic analysis: models and estimation procedures. *Am. J. Hum. Genet.* **19**, 233–257 (1967).
- Castonguay, M., Hodson, P. V., Moriarty, C., Drinkwater, K. F. & Jessop, B. M. Is there a role of ocean environment in American and European eel decline? *Fish. Oceanogr.* **3**, 197–203 (1994).
- Tsukamoto, K., Nakai, I. & Tesch, W.-V. Do all freshwater eels migrate? *Nature* **396**, 635–636 (1998).
- Raymond, M. & Rousset, F. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J. Hered.* **86**, 248–249.
- Rice, W. R. Analysing tables of statistical tests. *Evolution* **43**, 223–225 (1989).
- Palumbi, S. R. Genetic divergence, reproductive isolation, and marine speciation. *Annu. Rev. Ecol. Syst.* **25**, 547–572 (1994).
- Waples, R. S. Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species. *J. Hered.* **89**, 438–450 (1998).
- Hutchinson, D. W. & Templeton, A. R. Correlation of pairwise genetic and geographic distance measures: inferring the relative influences of gene flow and drift on the distribution of genetic variability. *Evolution* **53**, 1898–1914 (1999).
- Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**, 209–220 (1967).
- Saitou, N. & Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
- Avise, J. C. *et al.* The evolutionary genetic status of Icelandic eels. *Evolution* **44**, 1254–1262 (1990).
- Lecomte-Finiger, R. The early life of the European eel. *Nature* **370**, 424 (1994).

25. Marchese, P. J. Variability in the Gulf stream recirculation gyre. *J. Geophys. Res.* **104**, 29549–29560 (1999).
26. Maniatis, T., Fritsch, E. F. & Sambrook, J. *Molecular Cloning. A Laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1982).
27. Felsenstein, J. PHYLIP (Phylogeny inference package) manual, version 3.5c (Univ. Washington, Seattle, 1993).

Acknowledgements

We are extremely grateful to the following individuals and organizations for providing the samples used in this study: P. Prouzet, G. Adam, M.-N. de Casamajor, S. L. Jónsdóttir, E. Feunteun, P. Lambert, P. Dumont, C. Briand, R. Leconte, A. Crivelli, C. Gazeau, M.-F. Gazerque, D. Fatin, H. Wickström, A. Yahyaoui, C. Antunes, E. Ciccotti, A. Völlestad, B. Knights, H. Wilkens and the CEMAGREF. We would also like to thank J. Dodson, M. Castonguay, S. Rogers and J. McNeil for helpful comments on an earlier version of the manuscript. This work is a contribution to the programme of GIROQ.

Correspondence and requests for materials should be addressed to T.W. (e-mail: wirth_t@mpiib-berlin.mpg.de).

Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain

Christopher C. Pack & Richard T. Born

Department of Neurobiology, Harvard Medical School, 220 Longwood Avenue, Boston, Massachusetts 02115, USA

A critical step in the interpretation of the visual world is the integration of the various local motion signals generated by moving objects. This process is complicated by the fact that local velocity measurements can differ depending on contour orientation and spatial position. Specifically, any local motion detector can measure only the component of motion perpendicular to a contour that extends beyond its field of view^{1,2}. This “aperture problem”³ is particularly relevant to direction-selective neurons early in the visual pathways, where small receptive fields permit only a limited view of a moving object. Here we show that neurons in the middle temporal visual area (known as MT or V5) of the macaque brain reveal a dynamic solution to the aperture problem. MT neurons initially respond primarily to the component of motion perpendicular to a contour’s orientation, but over a period of approximately 60 ms the responses gradually shift to encode the true stimulus direction, regardless of orientation. We also report a behavioural correlate of these neural responses: the initial velocity of pursuit eye movements deviates in a direction perpendicular to local contour orientation, suggesting that the earliest neural responses influence the oculomotor response.

If a vertically orientated bar moves up and to the right at a constant velocity, small receptive fields positioned along the length of the contour can measure only the rightward component of motion, as the upward component provides no time-varying information (Fig. 1a). In contrast, cells positioned at the endpoints of the contour can measure motion direction accurately. Since direction-selective cells in the primary visual cortex (V1) have extremely small receptive fields, they are constantly faced with this aperture problem. Moreover, they provide directional input to subsequent stages of visual processing, which could perpetuate errors in motion computation. How are these conflicting motion signals ultimately resolved in the visual cortex? A candidate neural substrate for this computation is the middle temporal visual area (MT or V5), where neurons are known to integrate directional

responses from V1⁴, and are capable of computing motion direction for complex patterns^{5–7}.

We used the stimulus illustrated in Fig. 1b to measure neuronal responses in MT of alert macaque monkeys to moving contours at different orientations. Each stimulus consisted of a field of small white bars against a dark background. The size of the bar field was matched to each cell’s classical receptive field. The length of each bar was always 3°, significantly longer than corresponding receptive fields in V1, but smaller than the excitatory receptive fields in parafoveal and peripheral MT⁸. The use of multiple bars ensured that local motion signals from contours and contour endpoints stimulated the MT receptive fields at each instant. (Additional experiments using single long bars yielded results similar to those reported below, but the bar field had the advantage of providing stimulation that was evenly distributed across the receptive field and relatively constant over time.) On each trial, the angle between motion direction and contour orientation (ϕ in Fig. 1a) was 45°, 90° or 135°, and the stimulus moved in one of eight directions. To separate the selectivity of MT cells to static stimulus orientation^{9,10} from their directional responses, the stimulus remained stationary for 240 ms before moving. Because the motion direction and relative orientation, ϕ , both varied in intervals of 45°, the orientation did not predict the subsequent motion direction. The preferred direction (PD) for each cell was computed as a vector average of the stimulus direction weighted by the response to that direction. We recorded data from 60 MT cells from three hemispheres in two adult rhesus monkeys.

Figure 2a shows the results from one MT neuron. The earliest direction-selective responses, which occurred at a latency of approximately 70 ms after the onset of stimulus motion, showed a clear dependence on bar orientation relative to motion direction (ϕ). When the cell was stimulated with a field of bars moving perpendicular to their orientation ($\phi = 90^\circ$, red lines), the best responses were obtained for motion down and to the left (PD = 224°). For the $\phi = 45^\circ$ case (blue lines), the best response occurred for motion to the left (PD = 191°), and in the $\phi = 135^\circ$ case (green lines), the best response occurred for downward motion (PD = 266°). The effects of contour orientation on the directional responses were highly significant ($P < 0.001$, Watson–Williams test), with the peak response always occurring at the same oblique bar orientation (Fig. 2a). These early responses can best be described as encoding the component of stimulus motion perpendicular to bar orientation. We note that they are not responses to orientation

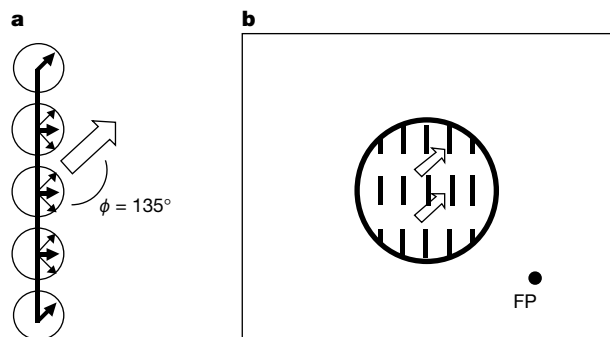


Figure 1 The aperture problem. **a**, Local motion detectors (indicated by the circles) along the contour can only measure motion perpendicular to the contour’s orientation. For these detectors, the direction of object motion is ambiguous because any of the physical velocities indicated by the thin black arrows would yield the same motion measurement (thick black arrows). The angle between contour orientation and motion direction, as measured clockwise from the motion direction (white arrow), is referred to as ϕ . **b**, The stimulus used in our experiments. A field of bars moved within a window sized to approximate the classical receptive field (depicted by the large circle) of each cell. FP, fixation point.

per se: motion up and to the right failed to excite the cell, even though the orientation was optimal. MT responses to static orientation are known to be quite small in amplitude, and so probably would not have directly influenced our results⁹. Figure 2b shows the direction tuning for the same cell when responses are averaged over a 1,500-ms period beginning 500 ms after the onset of stimulus motion. In this case, the best response is always obtained for motion down and to the left, with no statistically significant effect of bar orientation ($P > 0.2$, Watson–Williams test). Clearly the initial dependence on bar orientation (Fig. 2a) decreases during the course of prolonged stimulation.

To study the temporal evolution of the shift in MT response properties, we computed the PD for each MT cell in 15-ms bins starting from the earliest direction-tuned responses. These occurred, on average, at 75 ms after the onset of stimulus motion ($P < 0.05$, Rayleigh Z-test for deviation from circular uniformity). For each bin, the PD was measured and aligned

relative to the PD computed in the time-averaged $\phi = 90^\circ$ case. Figure 2c shows the average difference in PDs at each 15-ms bin for the population of 60 MT cells. The earliest responses are highly dependent on stimulus orientation, and this dependence decreases gradually over the course of approximately 60 ms. Within 150 ms after the onset of stimulus motion, MT cells primarily encode the actual stimulus direction, irrespective of orientation. However, there is a slight effect of stimulus orientation on the time-averaged relative PDs (computed in the interval from 500 ms to 2,000 ms after the onset of stimulus motion), resulting in a mean difference of $-3.1^\circ \pm 16^\circ$ for the $\phi = 45^\circ$ case and $5.3^\circ \pm 10.5^\circ$ for the $\phi = 135^\circ$ case. These small mean differences are nevertheless statistically significant ($P < 0.05$, paired one-tailed t -test), indicating a slight residual effect of bar orientation on the directional response.

For each cell in our MT population, we also examined the preferred orientation for static bar stimuli, as measured in the 240 ms before the onset of stimulus motion. We found these static bar responses to be quite weak and extremely variable. Moreover, we did not find any consistent relationship between the preferred orientation relative to preferred direction and the temporal properties of the directional response. This was unexpected in light of previous work demonstrating that such orientation preferences are predictive of MT responses to more complex “plaid” stimuli⁶. One possible explanation lies in the fact that previous studies were conducted on anaesthetized animals, whereas our experiments were conducted on awake, behaving animals. We have previously found that, in alert monkeys, the sudden appearance of a stimulus generates a strong response that is not dependent on stimulus features, such as orientation¹¹, and this may have weakened the selectivity for orientation. It is also possible that the orientation-selective inputs from V1 are not homogeneous across the MT receptive fields, so that our bar field stimulus activated inputs tuned to many different orientations.

There is strong evidence that MT neurons provide motion signals for the initiation of smooth pursuit eye movements^{12–14}. Because pursuit initiation typically occurs approximately 100 ms after the onset of a moving target¹⁵, the current results predict an orientation-dependent bias in initial eye velocity. To examine this possibility, we trained two monkeys to pursue the centre of a single orientated bar using the same parameters of motion direction and bar orientation as in the physiological recordings. Eye position was monitored with a scleral search coil¹⁶, and the resulting measurements were used to compute instantaneous eye velocity. We then aligned the eye velocity traces from each trial relative to the actual direction of stimulus motion, and computed the set of median traces across all trials (Fig. 3a). For accurate pursuit movements, eye velocities should match the direction and speed of the target, and indeed this occurs for the $\phi = 90^\circ$ case (red line). For the $\phi = 45^\circ$ and $\phi = 135^\circ$ cases (blue and green lines, respectively), the eye velocity deviates substantially in a direction perpendicular to the orientation of the bar. Note that the recovery of the eye movement from its initial deviation is somewhat slower than the recovery of the MT responses (Fig. 2c). This is likely to be due to differences in the stimuli, which were not equated for speed or bar length. In general the magnitude of the oculomotor effect scaled with bar length, such that the deviation of eye velocity increased monotonically as the size of the bar was varied from 5° to 25° . There was no consistent effect of orientation on eye velocity parallel to the actual direction of stimulus motion (Fig. 3b).

The neural effects of contour orientation on motion integration demonstrated here have a perceptual correlate in humans. It has been shown that the perceived direction of a field of moving bars is initially perpendicular to the orientation of the bars, and rotates towards the actual direction over a period of approximately 200 ms (ref. 17). This time course is somewhat slower than that found for our MT responses, but such a difference is to be expected on the

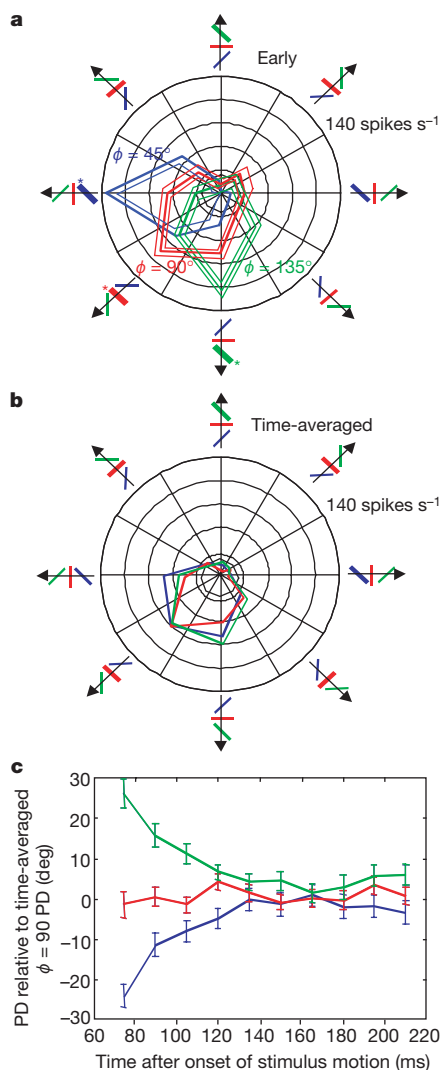


Figure 2 Evolution of direction tuning. **a**, Earliest direction-tuned response for a single MT neuron. Direction tuning is represented in polar coordinates with axes of stimulus direction (angle) and cell response in spikes per second (radius). Each tuning curve corresponds to a different value of ϕ , which is the angle between bar orientation and motion direction. Thin lines around each tuning curve represent standard error of the mean. **b**, Direction tuning for the same MT cell averaged over the last 1,500 ms of the stimulus presentation. **c**, Directional response as a function of time for the population of 60 MT cells. Error bars indicate standard error of the mean. PD, preferred direction (see text).

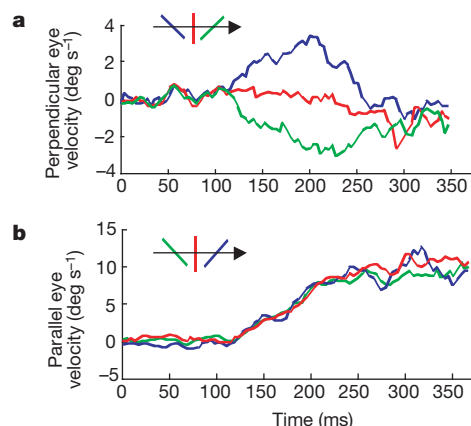


Figure 3 Visual tracking of an orientated bar. **a**, Median eye velocity perpendicular to the direction of target motion over time for the three bar orientations ($\phi = 45^\circ, 90^\circ$ and 135°). **b**, Median eye velocity parallel to the direction of target motion for the same experiments. Error bars (representing standard error of the mean) are smaller than the width of the data lines.

basis of variations in stimulus speed and contrast between the two experiments. For “type II” plaid stimuli, observers initially perceive a vector average of the local motion signals, with the percept subsequently shifting to an intersection of constraints¹⁸. For these stimuli there is also a residual bias in the perceived direction that is comparable to what we have observed in our time-averaged MT responses¹⁸. Similarly, human tracking eye movements in response to “barber’s pole” stimuli initially deviate in a direction perpendicular to the orientation of contours, and are subsequently affected by the geometry of the surrounding aperture^{19,20}. Taken together, these results suggest that the primate visual system derives an initial estimate of motion direction by integrating ambiguous and unambiguous local motion signals over a large spatial range (at least 20°), and refines this estimate over time. We have shown that this temporal evolution can be seen at the level of single MT neurons. However, it may also involve neural networks that propagate unambiguous signals to “fill in” the missing information at ambiguous retinal locations^{21,22}. Such networks have been the focus of recent neural models that use recurrent²³ or feedback²⁴ processes to compute motion direction. It is also possible that the visual system computes the motion of contours and contour endpoints via different pathways, with the latter requiring a slightly longer latency^{17,25,26}. The present identification of a neural signature for this process should open the door to future experiments to elucidate its mechanism. □

Methods

Animals were seated comfortably in a standard primate chair (Crist Instruments) with their heads fixed. They were required to fixate a small red square displayed on a computer monitor at a distance of 57 cm in order to obtain a liquid reward. For physiological experiments, the fixation point was always positioned so that the receptive field was at the centre of the monitor. Stimuli were displayed at a mean luminance of 0.552 cd m^{-2} , against a black background (luminance 0.025 cd m^{-2}). Microelectrode recordings were obtained from 60 single units in parafoveal and peripheral MT (mean receptive field eccentricity 11.1° , mean diameter 12.7°). For physiological experiments, the stimulus consisted of a field of moving bars. Each bar subtended 3° of visual angle, and was centred at a point along an invisible grid, placed approximately in the centre of the receptive field. The spacing between points in the grid was 5° in both the vertical and horizontal directions. The size of the bar field was approximately matched to that of the classical receptive field of each cell. Bars that moved outside the window disappeared, but were replaced with new bars so that the area of stimulation within the receptive field was approximately constant over time. The speed of stimulus motion was also matched to each cell’s preference, as measured with moving dots or bars.

During the physiological experiments, all different stimulus conditions (8 possible directions, 3 possible values of ϕ) were randomly interleaved. For the earliest directional plot (Fig. 2a), spikes were collected during the interval from 60 to 80 ms after the onset of stimulus motion, and averaged across 10 identical stimulus presentations. The later

responses (Fig. 2b) are averaged across the same 10 stimulus presentations, but include only the last 1,500 ms of the stimulus presentation. For the population data (Fig. 2c), the preferred direction is measured in 15-ms bins and aligned relative to the preferred direction computed in the time-averaged $\phi = 90^\circ$ case for each cell.

Pursuit experiments were conducted after physiology experiments. On each trial, a bar appeared centred on the fixation point and immediately started moving in one of eight directions at 10° s^{-1} . The bar was 20° in length, and had a small, dim, red, gaussian blob (0.23 cd m^{-2}) in the centre. Direction and bar orientation were randomly interleaved across trials, and the monkey had to pursue the centre of the bar to within an accuracy of 2° to receive a liquid reward. Eye movements were monitored using the scleral search coil technique¹⁶. Both eye position and the eye velocity (obtained by analog differentiation, d.c. to 50 Hz, $-20 \text{ dB per decade}$) were sampled at 1 kHz and stored to disk at 250 Hz for subsequent off-line analysis. To permit comparison across trials having different directions of target motion, the eye movement data were rotated relative to the target direction. Further details of animal preparation and data collection are given elsewhere²⁷.

Received 11 September; accepted 11 December 2000.

- Wallach, H. Über visuell wahrgenommene Bewegungsrichtung. *Psychol. Forsch.* **20**, 325–380 (1935).
- Wuerger, S., Shapley, R. & Rubin, N. “On the visually perceived direction of motion,” by Hans Wallach: 60 years later. *Perception* **11**, 1317–1367 (1996).
- Marr, D. & Ullman, S. Directional selectivity and its use in early visual processing. *Proc. R. Soc. Lond. B.* **211**, 151–180 (1981).
- Movshon, J. A. & Newsome, W. T. Visual response properties of striate cortical neurons projecting to area MT in macaque monkeys. *J. Neurosci.* **16**, 7733–7741 (1996).
- Movshon, J. A., Adelson, E. H., Gizzi, M. S. & Newsome, W. T. The analysis of moving visual patterns. *Exp. Brain Res. Suppl.* **11**, 117–151 (1986).
- Rodman, H. R. & Albright, T. D. Single-unit analysis of pattern-motion selective properties in the middle temporal visual area (MT). *Exp. Brain Res.* **75**, 53–64 (1989).
- Stoner, G. R. & Albright, T. D. Neural correlates of perceptual motion coherence. *Nature* **358**, 412–414 (1992).
- Albright, T. D. & Desimone, R. Local precision of visuotopic organization in the middle temporal area (MT) of the macaque. *Exp. Brain Res.* **65**, 582–592 (1987).
- Maunsell, J. H. & Van Essen, D. C. Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation. *J. Neurophysiol.* **49**, 1127–1147 (1983).
- Albright, T. D. Direction and orientation selectivity of neurons in visual area MT of the macaque. *J. Neurophysiol.* **52**, 1106–1130 (1984).
- Pack, C. C. & Born, R. T. Latency of direction tuning in cortical area MT of alert macaque. *Soc. Neurosci. Abstr.* **25**, 673 (1999).
- Newsome, W. T., Wurtz, R. H., Dürsteler, M. R. & Mikami, A. Deficits in visual motion processing following ibotenic acid lesions of the middle temporal visual area of the macaque monkey. *J. Neurosci.* **5**, 825–840 (1985).
- Groh, J. M., Born, R. T. & Newsome, W. T. How is a sensory map read out? Effects of microstimulation in visual area MT on saccades and smooth pursuit eye movements. *J. Neurosci.* **17**, 4312–4330 (1997).
- Lisberger, S. G. & Movshon, J. A. Visual motion analysis for pursuit eye movements in area MT of macaque monkeys. *J. Neurosci.* **19**, 2224–2246 (1999).
- Krauzlis, R. J. & Lisberger, S. G. Temporal properties of visual motion signals for the initiation of smooth pursuit eye movements in monkeys. *J. Neurophysiol.* **72**, 150–162 (1994).
- Robinson, D. A method of measuring eye movement using a scleral search coil in a magnetic field. *IEEE Trans. Biomed. Eng.* **10**, 137–145 (1963).
- Lorençau, J., Shiffrar, M., Wells, N. & Castet, E. Different motion sensitive units are involved in recovering the direction of moving lines. *Vision Res.* **33**, 1207–1217 (1993).
- Yo, C. & Wilson, H. R. Perceived direction of moving two-dimensional patterns depends on duration, contrast and eccentricity. *Vision Res.* **32**, 135–147 (1992).
- Masson, G. S., Rybarczyk, Y., Castet, E. & Mestre, D. R. Temporal dynamics of motion integration for the initiation of tracking eye movements at ultra-short latencies. *Vis. Neurosci.* **17**, 753–767 (2000).
- Beutter, B. R. & Stone, L. S. Human motion perception and smooth eye movements show similar directional biases for elongated apertures. *Vision Res.* **38**, 1273–1286 (1998).
- Hildreth, E. C. *The Measurement of Visual Motion* (MIT Press, Cambridge, Massachusetts, 1984).
- Watanabe, T. & Cole, R. Propagation of local motion correspondence. *Vision Res.* **35**, 2853–2861 (1995).
- Lidén, L. H. & Pack, C. C. The role of terminators and occlusion cues in motion integration and segmentation: A neural network model. *Vision Res.* **39**, 3301–3320 (1999).
- Chey, J., Grossberg, S. & Mingolla, E. Neural dynamics of motion grouping: From aperture ambiguity to object speed and direction. *J. Opt. Soc. Am. A* **14**, 2570–2594 (1997).
- Wilson, H. R., Ferrera, V. P. & Yo, C. A psychophysically motivated model for two-dimensional motion perception. *Vis. Neurosci.* **1**, 79–97 (1992).
- Duncan, R. O., Albright, T. D. & Stoner, G. R. Occlusion and the interpretation of visual motion: perceptual and neuronal effects of context. *J. Neurosci.* **20**, 5885–5897 (2000).
- Born, R. T., Groh, J. M., Zhao, R. & Lukaszewicz, S. J. Segregation of object and background motion in visual area MT: effects of microstimulation on eye movements. *Neuron* **26**, 725–734 (2000).

Acknowledgements

We thank P. Abrams for technical assistance, and J. Assad and T. Watanabe for comments on a previous version of the manuscript. This work was supported by a McDonnell-Pew Cognitive Neuroscience grant to C.C.P., and grants from NIH/NEI and The Giovanni Armenise-Harvard Foundation for Scientific Research to R.T.B.

Correspondence and requests for materials should be addressed to C.C.P. (e-mail: cpack@hms.harvard.edu).

A sodium-channel mutation causes isolated cardiac conduction disease

Hanno L. Tan^{*†‡}, Margreet T. E. Bink-Boelkens[§], Connie R. Bezzina^{||}, Prakash C. Viswanathan[†], Gertie C. M. Beaufort-Krol[§], Peter J. van Tintelen[¶], Maarten P. van den Berg[#], Arthur A. M. Wilde^{*} & Jeffrey R. Balser[†]

^{*} The Experimental and Molecular Cardiology Group, and ^{||} Department of Clinical Genetics, Academic Medical Center, University of Amsterdam, 1105 AZ Amsterdam, The Netherlands

[§] Department of Pediatric Cardiology, Beatrix Children's Hospital, 9700 RB Groningen, The Netherlands

[¶] Department of Medical Genetics, University Hospital, 9700 RB Groningen, The Netherlands

[#] Department of Cardiology, University Hospital, 9700 RB Groningen, The Netherlands

[†] Departments of Anesthesiology and Pharmacology, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA

[‡] These authors contributed equally to this work

Cardiac conduction disorders slow the heart rhythm and cause disability in millions of people worldwide. Inherited mutations in *SCN5A*, the gene encoding the human cardiac sodium (Na^+) channel, have been associated with rapid heart rhythms that occur suddenly and are life-threatening^{1–3}; however, a chief function of the Na^+ channel is to initiate cardiac impulse conduction. Here we provide the first functional characterization of an *SCN5A* mutation that causes a sustained, isolated conduction defect with pathological slowing of the cardiac rhythm. By analysing the *SCN5A* coding region, we have identified a single mutation in five affected family members; this mutation results in the

substitution of cysteine 514 for glycine (G514C) in the channel protein. Biophysical characterization of the mutant channel shows that there are abnormalities in voltage-dependent 'gating' behaviour that can be partially corrected by dexamethasone, consistent with the salutary effects of glucocorticoids on the clinical phenotype. Computational analysis predicts that the gating defects of G514C selectively slow myocardial conduction, but do not provoke the rapid cardiac arrhythmias associated previously with *SCN5A* mutations.

We have studied a family who came to medical attention when the proband, a 3-yr-old girl (pedigree: Fig. 1a, III-2), experienced episodes of fainting during a febrile illness. Her 12-lead electrocardiogram (ECG) showed characteristics of slow conduction throughout the atria and ventricles, including broad P waves, PR interval prolongation, and a wide QRS complex (Fig. 1c). Continuous heart rhythm monitoring (Fig. 1d) revealed episodes of severe bradycardia ($25 \text{ beats min}^{-1}$), which might arise from a reduced firing rate by the sinus node (native cardiac pacemaker), or from suppressed conduction through atrial tissues in the vicinity of the sinus node. During these slow periods, the cardiac rhythm was maintained by infrequent atrioventricular nodal 'escape' impulses (Fig. 1d, arrows).

The conduction disturbance persisted after the febrile illness resolved, and we determined no evidence of structural heart disease (echocardiography) or systemic diseases associated with conduction defects in children (viral infections, auto-immune or Lyme disease, thyroid dysfunction). Therapeutic intervention with a dual-chamber pacemaker was initially limited by inability to pace the atrium (maximal stimulus: 10 V, 1 ms); however, this difficulty resolved with 1 week of empirical steroid treatment (given originally to exclude an inflammatory aetiology for the conduction disturbance), which comprised intravenous methylprednisolone followed by oral prednisone.

During the 4 years following diagnosis, the proband has

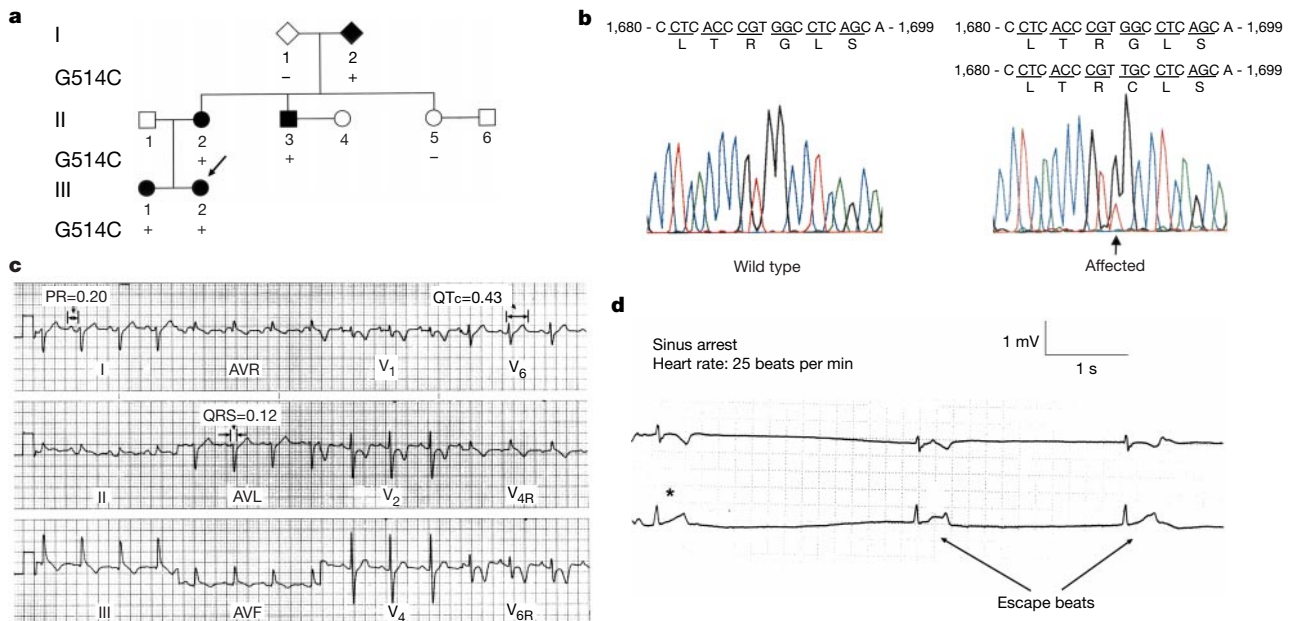


Figure 1 Genotype and ECG phenotype. **a**, Pedigree (arrow, proband) shows individuals consenting to SSCP analysis. +, carrier; –, non-carrier; square, male; circle, female; diamond, anonymous; filled symbol, conduction disease. **b**, Sequence identification of a G to T transversion (arrow) of one *SCN5A* allele in an affected individual (both alleles shown), causing replacement of glycine 514 by cysteine (G514C). **c**, Twelve-lead ECG of proband, indicating sinus rhythm with prolonged PR and QRS intervals (age-corrected normal values, 0.08–0.16 s and 0.03–0.07 s, respectively)²⁴ and a rightward QRS axis shift, but

no signs of Brugada or long QT syndrome. The P wave and QRS interval reflect atrial and ventricular conduction, respectively, and the PR interval reflects atrioventricular conduction. The QTc reflects ventricular repolarization, and is the QT interval (indicated on the figure) corrected for heart rate ($\text{QT}/\text{cycle length}^{0.5}$). **d**, Rhythm strip of proband shows typical bradycardia episode. The pauses after the last sinus beat (asterisk) are terminated by escape beats from a subsidiary pacemaker.

continuously required dual-chamber pacing. Examination of the family revealed nearly identical findings in a 6-yr-old sister (Fig. 1a, III-1). She also underwent pacemaker implantation with episodes of non-capture that reproducibly resolved with corticosteroid therapy. Three other family members (Fig. 1a, I-2, II-2, II-3) with no structural heart disease had ECG evidence of conduction slowing (prolonged PR and QRS intervals), but did not experience bradycardia or require pacemaker implantation. Single-strand conformation polymorphism (SSCP) analysis of the coding region of *SCN5A* identified an aberrant conformer in exon 12 in all five clinically affected family members. DNA sequencing in each of these individuals revealed a G to T transversion at the first nucleotide of codon 514 (Fig. 1b), resulting in the replacement of glycine by cysteine (G514C). The mutation was absent in 200 alleles from unaffected control individuals. In the proposed topology of the voltage-gated Na^+ channel⁴, this mutation is located intracellularly in the I–II interdomain linker.

To identify a functional mechanism for the slow-conducting phenotype, we expressed wild-type and G514C Na^+ channels in tsA201 cells to allow voltage-clamp measurements. On depolarization, Na^+ channels rapidly activate (open) and inactivate, producing a transient inward whole-cell current (I_{Na} ; Fig. 2a). G514C I_{Na} and wild-type I_{Na} were similar in size (Fig. 2, legend), suggesting that the mutation did not change expression; however, G514C I_{Na} decayed more rapidly than wild-type I_{Na} (Fig. 2a). The decay rate was bi-exponential (Fig. 2a), and the predominant, more rapid time

constant (τ_{fast}) was reduced by the mutation (Fig. 2b, –10 to 20 mV, $P < 0.05$), consistent with hastened inactivation of open channels. In addition, the magnitude of I_{Na} was shifted by +10 mV on the voltage axis (Fig. 2c, $P = 0.0003$), indicating a modified voltage dependence of activation. Both gating defects would reduce the I_{Na} available to support cardiac impulse conduction.

Further studies indicated that there were counterbalancing G514C gating effects that would increase the availability of I_{Na} (Figs 3 and 4a). Although mutant open-state inactivation was hastened (Fig. 2a, b), Na^+ channels also inactivate at membrane potentials below the opening threshold (so-called ‘closed-state’ inactivation)⁵. Examination of steady-state G514C availability at potentials negative to the opening threshold (Fig. 3a) revealed a +7 mV shift ($P = 0.003$ versus wild type), suggesting that closed-state inactivation was reduced.

A more detailed kinetic analysis using a variable-duration pre-pulse (Fig. 3b) found no significant change in the inactivation time constant (–90 mV: $\tau = 75 \pm 5$ ms for G514C versus 69 ± 8 ms for wild type; not significant), but the extent of G514C closed-state inactivation was reduced greatly (Fig. 3b, legend). Hence, the results in Fig. 3 suggest that the mutant inactivated state (once occupied)

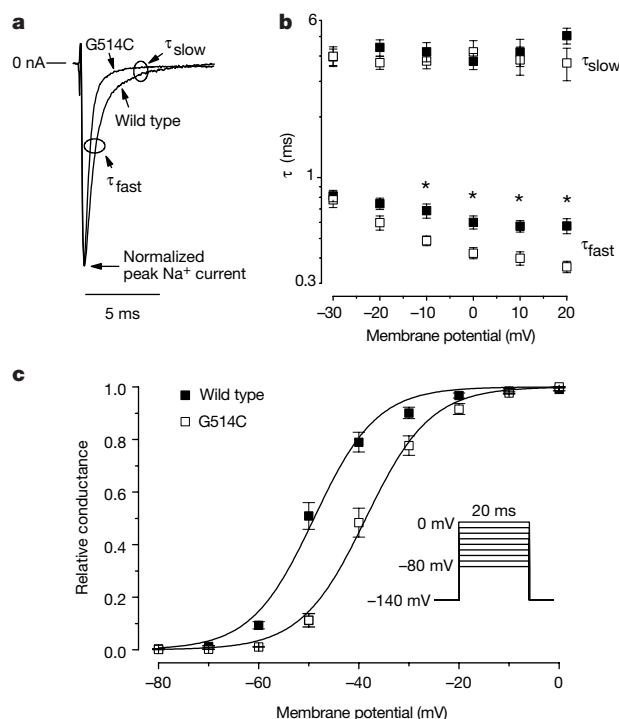


Figure 2 G514C gating effects that reduce I_{Na} availability. **a**, Wild-type and G514C I_{Na} transients are normalized to illustrate differences in decay rate. Peak I_{Na} density at 0 mV: wild-type ($n = 6$), 0.76 ± 0.10 nA pF⁻¹; G514C ($n = 7$), 0.56 ± 0.08 nA pF⁻¹ (not significant). **b**, I_{Na} decay was fitted by the function $y = A_1(1 - \exp[-t/\tau_{\text{fast}}]) + A_2(1 - \exp[-t/\tau_{\text{slow}}])$. τ_{fast} , τ_{slow} are plotted for wild type ($n = 6$) and G514C ($n = 7$) ($*P < 0.05$ versus wild type). At –10 mV: wild type, $A_1 = 0.89 \pm 0.03$, $A_2 = 0.11 \pm 0.03$; G514C, $A_1 = 0.94 \pm 0.01$, $A_2 = 0.06 \pm 0.01$ (not significant). **c**, Voltage-dependence of activation. Relative conductance was fitted to a Boltzmann function $y = [1 + \exp\{(V - V_{1/2})/k_B\}]^{-1}$. Wild type ($n = 6$) $V_{1/2} = -48.6 \pm 1.2$ mV, $k_B = 6.1 \pm 0.8$; G514C ($n = 7$) $V_{1/2} = -38.5 \pm 1.4$ mV ($P = 0.0003$ versus wild type), $k_B = 6.1 \pm 0.3$. At 32 °C: G514C ($n = 5$; not shown) $V_{1/2} = -37.9 \pm 3.2$ mV (not significant versus 22 °C), $k_B = 5.6 \pm 0.9$ (not significant versus 22 °C).

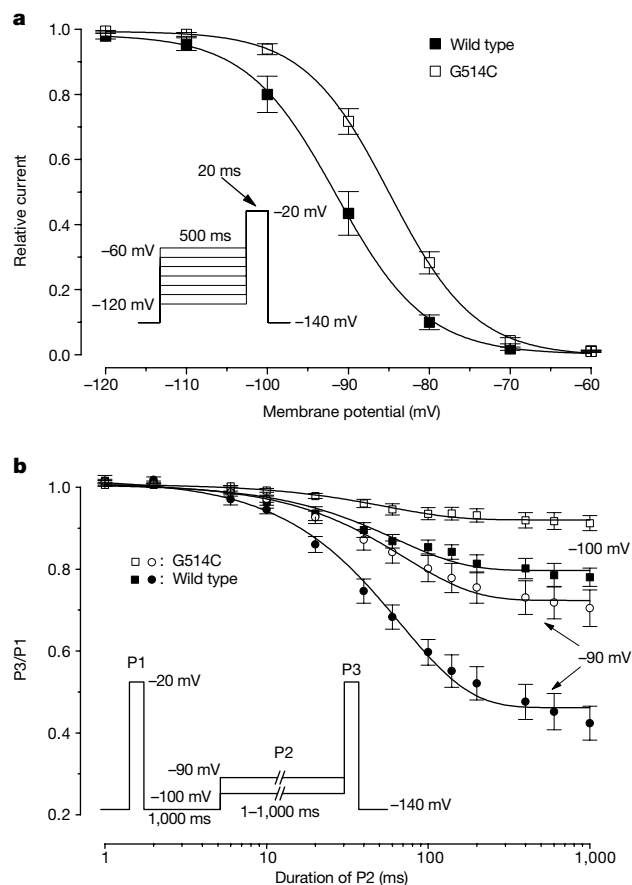


Figure 3 G514C gating effects that increase I_{Na} . **a**, The voltage dependence of steady-state availability (protocol shown, see Methods), fitted by a Boltzmann function (see Fig. 2c). Wild type ($n = 6$) $V_{1/2} = -92.0 \pm 1.7$ mV, $k_B = 5.4 \pm 0.3$; G514C ($n = 7$) $V_{1/2} = -85.1 \pm 0.9$ ($P = 0.003$ versus wild type), $k_B = 5.2 \pm 0.3$. At 32 °C: G514C ($n = 5$; not shown) $V_{1/2} = -84.1 \pm 2.1$ (not significant versus 22 °C), $k_B = 5.5 \pm 0.4$ (not significant versus 22 °C). **b**, Closed-state inactivation at membrane potentials near V_{rest} (protocol shown). Solid lines indicate least-squares fits to the function $y = y_0 + A \exp(-t/\tau)$. At –90 mV: wild type ($n = 6$) $\tau = 69 \pm 8$ ms, $A = 0.56 \pm 0.05$; G514C ($n = 7$) $\tau = 75 \pm 5$ ms (not significant versus wild type), $A = 0.28 \pm 0.04$ ($P < 0.01$ versus wild type). At –100 mV: wild type $\tau = 70 \pm 13$ ms, $A = 0.21 \pm 0.02$; G514C $\tau = 56 \pm 6$ ms (not significant versus wild type), $A = 0.08 \pm 0.01$ ($P < 0.001$ versus wild type).

was partly destabilized. Fig. 4a shows speeded recovery of inactivated G514C channels during hyperpolarization, further indicating that the mutant inactivated state was destabilized. Bath-applied reduced glutathione (1 mM, $n = 7$) did not modify G514C gating (data not shown), suggesting that these diverse functional defects (Figs 2–4) did not result from cysteine oxidation, but rather from allosteric effects caused by residue substitution that involve several gating domains.

Given the clinical improvement of two patients receiving steroid therapy, we analysed the effects of steroids on G514C gating. We incubated G514C and wild-type cells with dexamethasone (1 μ M) for 48 h after transfection. Dexamethasone partially rescued the mutation-induced shift in the voltage dependence of activation (Fig. 4b, left), but did not change G514C inactivation (Fig. 4b, right) and had no effect on wild-type gating (Fig. 4b). There was no effect of steroid pretreatment on I_{Na} density (Fig. 4b, legend) or on the rate of I_{Na} decay. Acute exposure of G514C to 1 μ M dexamethasone (≤ 30 min) through either the intracellular (pipette) or extracellular solution had no effects ($n = 6$ in each case; data not shown),

suggesting that the selective effect on G514C activation gating did not result from a direct interaction between dexamethasone and the Na^+ channel. The method whereby steroids influenced G514C gating remains speculative, but the requirement for sustained exposure suggests that the mechanisms involve channel synthesis, post-translational modification, or modified expression of auxiliary subunits.

Studies of Na^+ -channel function suggest that Brugada syndrome mutations reduce I_{Na} , hastening epicardial repolarization and causing idiopathic ventricular fibrillation^{6–8}, whereas mutations that evoke the long QT syndrome (LQT3) cause excessive I_{Na} , delaying repolarization and causing a distinctive ventricular tachyarrhythmia (Torsades de Pointes)^{1,2,9}. G514C evokes gating effects reminiscent of both syndromes (an activation gating shift that tends to reduce I_{Na} , and destabilized inactivation that tends to increase I_{Na}), but G514C carriers do not exhibit characteristic features of either LQT3 or Brugada syndrome (Fig. 1c). We thought that the ‘balance’ of the opposing G514C gating effects may produce isolated conduction slowing without evoking the Brugada syndrome or LQT3 repolarization defects.

We used computational studies to examine a model fibre that recapitulates layer-specific cardiac action-potential characteristics and cell-to-cell conduction¹⁰. First, we examined whether the G514C gating defects could reproduce the clinically observed conduction delay. Experimental studies^{11,12} indicate that increasing the voltage difference (ΔV) between the cardiac resting potential (V_{rest}) and the Na^+ -channel activation (opening) threshold decreases conduction velocity. We explored the isolated effect of reducing ΔV by shifting the voltage dependence of activation (as in G514C; Fig. 2c), rather than by depolarizing V_{rest} (with external K^+ in the experimental studies)^{11,12}. The model simulations (Fig. 5a) show a striking decrease in conduction velocity (left ordinate) as activation gating is shifted. There is also a marked increase in the minimum voltage stimulus necessary to support conduction through the fibre (right ordinate). The second observation might provide a rationale for the relative inexcitability of G514C carriers to pacemaker stimulation, and might also link the dexamethasone effects on G514C activation (Fig. 4b) to the improved pacemaker response with steroid therapy.

Given that these simulations linked the G514C activation gating defect to conduction slowing, we considered how the additional inactivation gating defect (Fig. 3a) would influence the model fibre. Epicardial cells exhibit a potassium current (I_{to}) that inscribes a ‘notch’ in the early plateau phase of the action potential (Fig. 5b, left). This repolarizing force renders the action potential ‘sensitive’ to factors that reduce I_{Na} (ref. 13). As such, Na^+ -channel mutations that reduce I_{Na} (that is, Brugada syndrome T1620M)³ severely shorten the action potential in the epicardium, but not in the inner-cell layers (Fig. 5b, middle). This layer-specific effect evokes a gradient in membrane potential from the epicardial to inner-cell layers, and may cause transmural current flow sufficient to induce the ECG changes and ventricular arrhythmias of Brugada syndrome^{14,15}. Although G514C also exhibits gating defects that reduce I_{Na} (Fig. 2), the mutation destabilizes closed-state inactivation at the resting membrane potential (Fig. 3a, b), an effect that should increase Na^+ -channel availability. Incorporating these

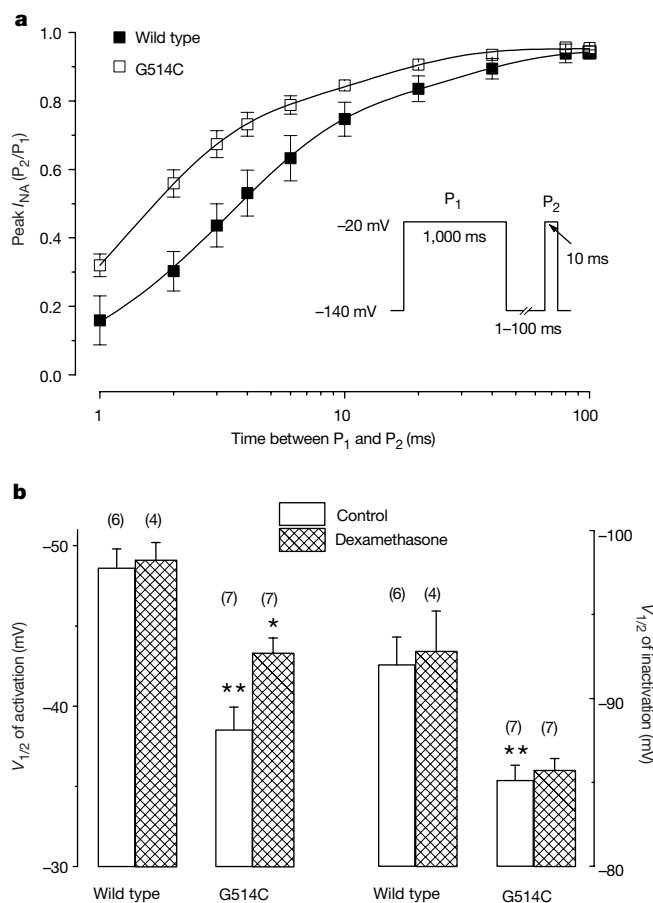


Figure 4 G514C inactivated-state instability, and steroid effects on gating. **a**, Recovery from inactivation (protocol shown, see Methods), fitted by the function $y = y_0 + A_1(1 - \exp[-t/\tau_1]) + A_2(1 - \exp[-t/\tau_2])$. Wild type ($n = 5$) $\tau_1 = 4.5 \pm 0.2$ ms, $\tau_2 = 81.7 \pm 17.3$ ms, $A_1 = 0.82 \pm 0.04$, $A_2 = 0.18 \pm 0.03$; G514C ($n = 7$) $\tau_1 = 2.1 \pm 0.4$, $\tau_2 = 48.0 \pm 6.2$, $A_1 = 0.84 \pm 0.01$, $A_2 = 0.15 \pm 0.01$ (τ_1 , $P = 0.007$ versus wild type). **b**, Effects of dexamethasone on G514C gating. Plotted are $V_{1/2}$ for activation (left ordinate) and inactivation (right ordinate) of wild type and G514C in the absence and presence of dexamethasone. G514C caused depolarizing shifts (** $P < 0.05$ versus wild type) in $V_{1/2}$ of activation and inactivation. Dexamethasone partially reversed the activation effect (* $P < 0.05$), but did not change inactivation gating or I_{Na} density (0 mV) of G514C (control, 0.56 ± 0.08 nA P_1^{-1} , $n = 7$; dexamethasone, 0.46 ± 0.05 , $n = 7$; not significant) or wild type (control, 0.76 ± 0.10 , $n = 6$; dexamethasone, 0.62 ± 0.05 , $n = 4$; not significant).

Table 1 Transverse conduction velocity

	Wild-type	Brugada syndrome (T1620M)	Conduction disease (G514C)	G514C + dexamethasone
Endocardial	22.8	18.8 (18%)	20.0 (12%)	21.0 (8%)
Midmyocardial	21.4	17.0 (21%)	18.4 (14%)	19.5 (9%)
Epicardial	20.4	15.8 (23%)	17.3 (15%)	18.4 (10%)

Calculated transverse conduction velocities ($cm s^{-1}$) between 10 cells in epicardial, endocardial and midmyocardial regions. The percentage change from wild-type value is given in parentheses. Simulation conditions were those described in Fig. 5b, including the heterozygous state for T1620M and G514C.

opposing gating effects into the model restored normal epicardial action-potential duration (Fig. 5b, right) while slowed conduction velocity persisted (Table 1), consistent with the clinical phenotype.

An activation gating defect similar to G514C (Fig. 2c) has been identified in T1620M⁶, consistent with clinical reports indicating that Brugada syndrome patients have marked abnormalities in ventricular conduction^{16,17}. In fact, the modelling results predict that destabilized closed-state inactivation gating of G514C may attenuate the ventricular conduction delay (Table 1, compare G514C with T1620M). In the three adult G514C carriers, the QRS intervals were 120 ms (Fig. 1a, pedigree I-2), 106 ms (II-2) and 110 ms (II-3)—values at the upper limit of normal—whereas the QRS intervals in Brugada syndrome patients are generally more prolonged. In the original report describing Brugada syndrome¹⁶, the QRS intervals were well above 120 ms in 6 out of 8 patients, and were 120 ms in the other 2 patients.

Analysis of the ECGs of G514C carriers indicates that steroid therapy improves not only atrial, but also ventricular conduction. In both children (Fig. 1a, III-1, III-2), the QRS duration during spontaneously conducting (non-paced) periods was shorter (by ~40 ms) after steroid treatment. Moreover, incorporating the antagonistic effect of dexamethasone on the G514C-induced activation shift (Fig. 4b) enhanced, but did not fully correct, the model fibre conduction velocity (Table 1) as seen clinically. Conduction velocity in the myocardium is heavily determined by the maximum

rate of rise of the action potential upstroke (Fig. 5b, boxed area on wild-type and G514C action potentials). This portion of the action potential is redrawn (Fig. 5b, inset) using an expanded time base to provide intuitive insight into how the activation gating effects of G514C and dexamethasone alter the conduction velocity in the model fibre. For the simulations shown, epicardial $dV/dt_{\max}$ was 306 V s^{-1} for wild type, but was only 242 V s^{-1} for G514C. The reduced activation shift associated with dexamethasone treatment of G514C yielded an intermediate $dV/dt_{\max}$ value (279 V s^{-1}), consistent with the model-predicted conduction velocities (Table 1).

The number of gating defects associated with the G514C mutation provides a biophysical mechanism for the conduction defects observed in this family. At the same time, *SCN5A* mutations have been identified in two families with inherited conduction disease that might, on the basis of rather large deleted segments, produce entirely non-functional Na^+ channels¹⁸. Although functional data are not available for these mutations, the lack of a tachyarrhythmia phenotype indicates that 'modifier' genes, allele penetrance and/or developmental factors may also influence the relationship between conduction disease and Na^+ -channel function. Nonetheless, the G514C channel provides new evidence that Na^+ -channel gating dysfunction can produce isolated conduction disease, with pathological slowing of the heart rhythm. This inherited lesion suggests a biophysical framework for understanding how the diverse functional behaviour of a single ion channel may evoke a spectrum of

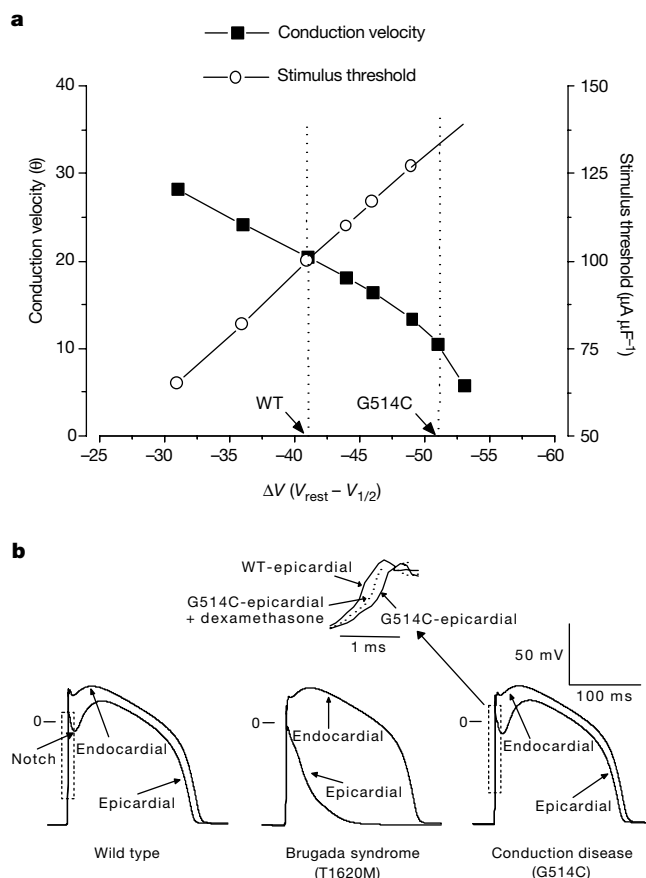


Figure 5 Action potential propagation in a model fibre. **a**, Model-predicted values for conduction velocity and stimulus magnitude required to support conduction, as a function of voltage dependence of activation. ΔV , difference (in mV) between activation $V_{1/2}$ and V_{rest} (about -90 mV). Dotted lines indicate ΔV for wild-type (WT) and G514C channels. **b**, Simulated endocardial and epicardial action potentials. Middle and Right panels simulate the heterozygous condition (50:50 mixture of mutant and wild-type channels). Left, wild-type action potentials; the epicardial 'notch' due to I_{to} is indicated.

Middle, T1620M; $V_{1/2}$ of activation shifted $+10 \text{ mV}$ and I_{Na} decay hastened to one-half of wild-type value (ref. 6). Right, G514C; $V_{1/2}$ of activation shifted $+10 \text{ mV}$ (Fig. 2c), I_{Na} decay speeded to two-thirds of wild-type value (Fig. 2a), and $V_{1/2}$ of inactivation shifted $+7 \text{ mV}$ (Fig. 3a). **b**, Inset, action potential upstroke (boxed regions on wild-type and G514C action potentials) shown on an expanded time base to illustrate $dV/dt_{\max}$. Dotted line indicates predicted G514C $dV/dt_{\max}$ in dexamethasone, where the G514C activation $V_{1/2}$ shift is reduced from $+10 \text{ mV}$ to $+5 \text{ mV}$ (Fig. 4b).

deranged cardiac excitability, from conduction disease to sudden cardiac death. □

Methods

Sequence analysis, mutagenesis and transfection

The entire coding region of *SCN5A* was screened by SSCP and direct sequence analysis as described¹⁹, and site-directed mutagenesis was performed on *SCN5A* complementary DNA cloned into the green fluorescent protein (GFP) expression vector pCGI (ref. 20) as described²¹. Cultured cells (tsA201) were co-transfected with a 5:1 molar ratio of the Na⁺-channel β_1 subunit (provided by A. George, Vanderbilt University).

Electrophysiology

Unless otherwise indicated, I_{Na} was recorded at room temperature (22 °C) to optimize voltage-clamp control using established recording conditions⁷. For all voltage-clamp experiments, the pipette solution contained (in mM): NaF 10, CsF 110, CsCl 20, EGTA 10 and HEPES 10 (pH 7.35 with CsOH), and the bath solution contained: NaCl 145, KCl 4, CaCl₂ 1.8, MgCl₂ 1, HEPES-NaOH pH 7.35. For voltage-dependent activation (Fig. 2c), relative conductance was calculated by dividing the peak I_{Na} at each clamp potential by the driving force (clamp voltage minus measured reversal potential), then normalizing to the value at 0 mV. For voltage-dependent inactivation (Fig. 3a), relative currents were calculated by normalizing peak I_{Na} (elicited by depolarization to -20 mV) to the value recorded after holding at the most negative (-120 mV) conditioning voltage. The rate of development of closed-state inactivation (Fig. 3b) was analysed using variable duration pre-pulses to -90 or -100 mV. The kinetics of recovery from inactivation (Fig. 4a) were assessed using a three-pulse clamp protocol. The ratio of the peak I_{Na} elicited by the P₂ pulse, relative to the P₁ pulse, was plotted as a function of the intervening recovery interval at -140 mV.

Action potential simulations

The Luo-Rudy model^{10,22,23} of a mammalian ventricular myocyte action potential provided the basis for the action potential simulations (Fig. 5), with I_{to} added as described⁶. A one-dimensional fibre¹⁰ consisting of endocardial (cells 1–80), midmyocardial (81–110), and epicardial (111–190) cells was used to mimic *in vivo* conditions. A stimulus was applied to cell no. 1 to simulate normal endocardial to epicardial impulse propagation.

Received 22 November; accepted 21 December 2000.

- Wang, Q. *et al.* *SCN5A* mutations associated with an inherited cardiac arrhythmia, long QT syndrome. *Cell* **80**, 805–811 (1995).
- Bennett, P. B., Yazawa, K., Naomasa, M. & George, A. L. Molecular mechanism for an inherited cardiac arrhythmia. *Nature* **376**, 683–685 (1995).
- Chen, Q. *et al.* Genetic basis and molecular mechanism for idiopathic ventricular fibrillation. *Nature* **392**, 293–296 (1998).
- Noda, M. *et al.* Expression of functional sodium channels from cloned cDNA. *Nature* **322**, 826–828 (1986).
- Horn, R., Patlak, J. B. & Stevens, C. F. Sodium channels need not open before they inactivate. *Nature* **291**, 426–427 (1981).
- Dumaine, R. *et al.* Ionic mechanisms responsible for the electrocardiographic phenotype of the Brugada syndrome are temperature dependent. *Circ. Res.* **85**, 803–809 (1999).
- Veldkamp, M. W. *et al.* Two distinct congenital arrhythmias evoked by a multidysfunctional Na⁺ channel. *Circ. Res.* **86**, E91–E97 (2000).
- Deschenes, I. *et al.* Electrophysiological characterization of *SCN5A* mutations causing long QT (E1784K) and Brugada (R1512W and R1432G) syndromes. *Cardiovasc. Res.* **46**, 55–65 (2000).
- Priori, S. G. *et al.* Genetic and molecular basis of cardiac arrhythmias: impact on clinical management parts I and II. *Circulation* **99**, 518–528 (1999).
- Viswanathan, P. C., Shaw, R. M. & Rudy, Y. Effects of IKr and IKs heterogeneity on action potential duration and its rate dependence: a simulation study. *Circulation* **99**, 2466–2474 (1999).
- Dominguez, G. & Fozzard, H. A. Influence of extracellular K⁺ concentration on cable properties and excitability of sheep cardiac Purkinje fibers. *Circ. Res.* **26**, 565–574 (1970).
- Casico, W. E., Johnson, T. A. & Gettes, L. S. Electrophysiologic changes in ischemic ventricular myocardium: I. Influence of ionic, metabolic, and energetic changes. *J. Cardiovasc. Electrophysiol.* **6**, 1039–1062 (1995).
- Krishnan, S. C. & Antzelevitch, C. Sodium channel block produces opposite electrophysiological effects in canine ventricular epicardium and endocardium. *Circ. Res.* **69**, 277–291 (1991).
- Yan, G. X. & Antzelevitch, C. Cellular basis for the Brugada syndrome and other mechanisms of arrhythmogenesis associated with ST-segment elevation. *Circulation* **100**, 1660–1666 (1999).
- Alings, M. & Wilde, A. “Brugada” syndrome: clinical data and suggested pathophysiological mechanism. *Circulation* **99**, 666–673 (1999).
- Brugada, P. & Brugada, J. Right bundle branch block, persistent ST segment elevation and sudden cardiac death: a distinct clinical and electrocardiographic syndrome. A multicenter report. *J. Am. Coll. Cardiol.* **20**, 1391–1396 (1992).
- Priori, S. G., Napolitano, C., Giordano, U., Collisani, G. & Memmi, M. Brugada syndrome and sudden cardiac death in children. *Lancet* **355**, 808–809 (2000).
- Schott, J. J. *et al.* Cardiac conduction defects associate with mutations in *SCN5A*. *Nature Genet.* **23**, 20–21 (1999).
- Bezzina, C. *et al.* A single Na⁺ channel mutation causing both long-QT and Brugada syndromes. *Circ. Res.* **85**, 1206–1213 (1999).
- Johns, D. C., Nuss, H. B. & Marban, E. Suppression of neuronal and cardiac transient outward currents by viral gene transfer of dominant negative KV4.2 constructs. *J. Biol. Chem.* **272**, 31598–31603 (1997).

- Kambouris, N. G. *et al.* Phenotypic characterization of a novel long QT syndrome mutation in the cardiac sodium channel. *Circulation* **97**, 640–644 (1998).
- Luo, C. H. & Rudy, Y. A dynamic model of the cardiac ventricular action potential. I. Simulations of ionic currents and concentration changes. *Circulation Research* **74**, 1071–1096 (1994).
- Zeng, J., Laurita, K. R., Rosenbaum, D. S. & Rudy, Y. Two components of the delayed rectifier K⁺ current in ventricular myocytes of the guinea pig type. Theoretical formulation and their role in repolarization. *Circ. Res.* **77**, 140–152 (1995).
- Davignan, A. *et al.* Normal ECG standards for infants and children. *Pediatr. Cardiol.* **1**, 123–152 (1979).

Acknowledgements

Financial support for this study was provided by the Interuniversity Cardiology Institute of Netherlands (H.L.T. and A.A.M.W.), the Dutch Heart Foundation (A.A.M.W.) and the NIH (J.R.B.). We thank S. Stepanovic and M. Hulsbeek for technical assistance, and L. DeFelice, D. Roden, M. Anderson and A. George for helpful discussion.

Correspondence and requests for materials should be addressed to J.R.B. (e-mail: jeff.balser@mcm.vanderbilt.edu).

The voltage-sensitive sodium channel is a bell-shaped molecule with several cavities

Chikara Sato*, Yutaka Ueno*, Kiyoshi Asai*, Katsutoshi Takahashi†, Masahiko Sato‡, Andreas Engel§ & Yoshinori Fujiyoshi||

* Supermolecular Science Division, Electrotechnical Laboratory (ETL), Umezono 1-1-4, Tsukuba, 305-8568 Japan

† School of Knowledge Science, Japan Advanced Institute of Science and Technology Hokuriku (JAIST), Asahidai 1-1, Tatsunokuchi, Ishikawa, 923-1211, Japan

‡ Central Research Institute, Itoham Foods Inc., Kubogaoka 1-2, Moriya, 302-0104, Japan

§ Maurice E. Müller Institute, at the Biozentrum, University of Basel, Klingelbergstrasse 70, CH-4056 Basel, Switzerland

|| Department of Biophysics, Faculty of Science, Kyoto University, Oiwake, Kitashirakawa, Sakyo-ku, Kyoto, 606-8502, Japan

Voltage-sensitive membrane channels, the sodium channel, the potassium channel and the calcium channel operate together to amplify, transmit and generate electric pulses in higher forms of life. Sodium and calcium channels are involved in cell excitation, neuronal transmission, muscle contraction and many functions that relate directly to human diseases^{1–4}. Sodium channels—glycosylated proteins with a relative molecular mass of about 300,000 (ref. 5)—are responsible for signal transduction and amplification, and are chief targets of anaesthetic drugs⁶ and neurotoxins¹. Here we present the three-dimensional structure of the voltage-sensitive sodium channel from the eel *Electrophorus electricus*. The 19 Å structure was determined by helium-cooled cryo-electron microscopy and single-particle image analysis of the solubilized sodium channel. The channel has a bell-shaped outer surface of 135 Å in height and 100 Å in side length at the square-shaped bottom, and a spherical top with a diameter of 65 Å. Several inner cavities are connected to four small holes and eight orifices close to the extracellular and cytoplasmic membrane surfaces. Homologous voltage-sensitive calcium and tetrameric potassium channels, which regulate secretory processes and the membrane potential⁷, may possess a related structure.

Different approaches have provided information on the structure of the sodium channel. Its function and folding topology have been assessed by site-directed mutagenesis^{8,9} and peptide-specific antibody binding^{10,11}. The structure of a synthetic peptide

corresponding to the III–IV linker (2.6% of the entire molecule) has been analysed by NMR¹², whereas the main features of the solubilized sodium channel have been determined by negative-stain electron microscopy¹³. The structure of a bacterial potassium channel KcsA mutant, which has 20% of the carboxy terminus deleted, has been solved to a resolution of 3.2 Å by MacKinnon's group¹⁴. The structure analysis based on X-ray crystallography provided a clear explanation for the ion selectivity of the potassium channel¹⁴; however, the structure of the entire molecule, which is predicted to have six transmembrane helices per subunit^{15–17}, is essential to understand the mechanism of voltage gating.

Sodium channel molecules were solubilized from *Electrophorus electricus* *electroplax*, immunoaffinity purified¹³ and imaged by cryo-electron microscopy¹⁸. The images had a poor signal to noise (S/N) ratio (Fig. 1), because the low contrast inherent to proteins was further decreased by the presence of 20% glycerol and 2 M MgCl₂, which were required to stabilize the detergent-solubilized sodium channel. The S/N ratio was improved by aligning and averaging projections of similarly orientated molecules (Fig. 1, row 2). Depending on their orientation in the frozen buffer layer, the sodium channels appeared as either almost-square or wedge-shaped projections. The square projections show a faint central

density with a central depression, as well as four massive densities at their corners (Fig. 1a, rows 1 and 2). The wedge-shaped projections are more variable, including various densities at peripheral and central regions (Fig. 1b–d, rows 1 and 2).

We interactively selected particles discernible above the background, and processed them for the three-dimensional (3D) reconstruction using the Imagic V system¹⁹. A total of 11,991 particles were aligned rotationally and translationally by the reference-free method²⁰, classified into 240 clusters and averaged. The averages were then refined in several cycles²¹. Because the averaged projection down the channel axis exhibited a pseudo-four-fold symmetry (Fig. 1a, row 2), the initial Euler angles of the cluster averages were determined²² assuming four-fold symmetry. Subsequently, the structure was refined without any symmetry constraints. The Euler angles of the 240 cluster averages (Fig. 1e) showed the sodium channels to be almost randomly orientated in the frozen buffer layer. Comparison of two independent reconstructions by the Fourier shell correlation documents a resolution of 19 Å (Fig. 1f).

The surface representations in Fig. 1, row 3, display the outer shapes of the protein at viewing angles defined by the Euler angles (noted below each column) of the cluster averages in Fig. 1, row 2. Projections along these directions calculated from the 3D density

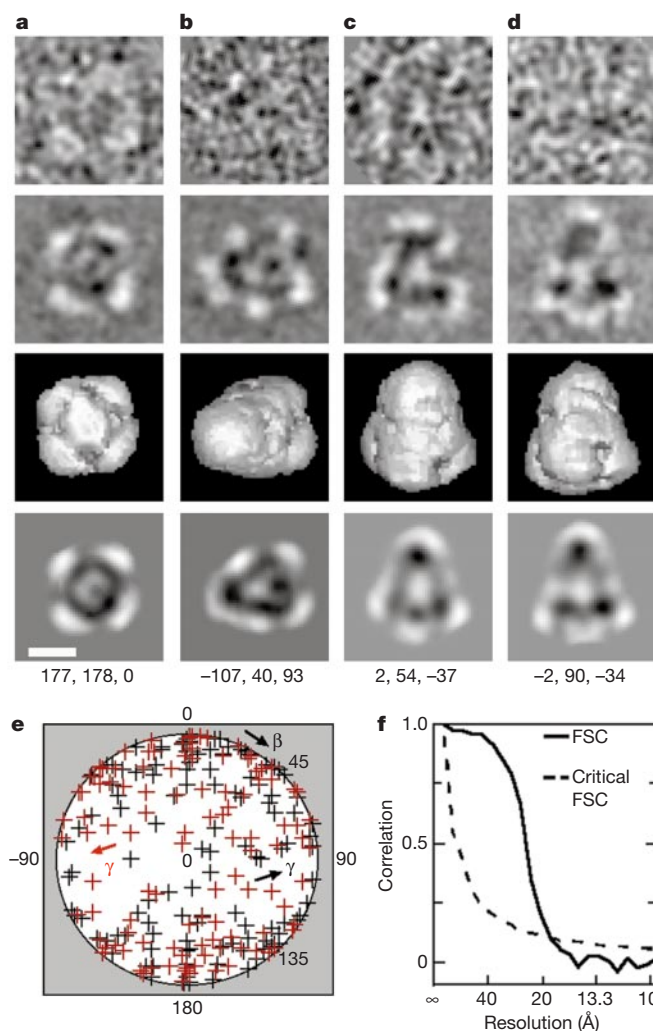


Figure 1 Cryo-electron microscopy of the sodium channel. **a–d**, Raw images of molecules with different Euler angles (row 1) are compared with the corresponding 2D averages (row 2), the surface views of the 3D reconstruction (row 3) and the projections of the 3D reconstruction (row 4) along the corresponding Euler directions. Protein is displayed in bright shades. Scale bar, 50 Å. **e**, Surface projection of the Euler angle

distribution of the 240 class averages obtained from 11,991 images. Each class average is represented by a cross in a (β , γ) coordinate system. **f**, Fourier shell correlation function between two reconstructions calculated from two half-sets of the data. The cross-over indicates a resolution limit of 19 Å.

map (Fig. 1, row 4) compare well with the experimental averages (Fig. 1, row 2). The isosurface in Fig. 1, row 3, encompasses 150% of the relative molecular mass (M_r) of the channel protein, 208,000 (208K), discounting the glycan moiety that is lost in the averages owing to disorder. At this contouring threshold, four small and four prominent holes penetrate the exterior surface (Fig. 1a–d, row 3). A higher threshold provides a structure with a more complex surface that is difficult to interpret. Examination of the 3D structure over the whole range of Euler angles (Fig. 2) reveals a bell-shaped molecule with a square-shaped end and a narrow top that resembles a half-sphere.

Longitudinal views of the molecule (Figs 1b–d, row 3; and 2) show a 30 Å wide horizontal band without peripheral holes, which probably represents the transmembrane domains (Fig. 2, white lines). Thus, the molecule consists of an upper narrower region corresponding to 24% of the total volume, a transmembrane region (29%), and a lower wider region comprising 47% of the total volume. The relative volumes were estimated using the isosurface shown in Fig. 2. The values for upper and lower regions are in good agreement with those expected from sequence prediction of the extracellular and cytoplasmic domains of the channel molecule (Fig. 3a)^{15–17}. Therefore, we tentatively assume that the upper region with its continuous half-spherical outer shell is extracellular, and that the lower region that has prominent orifices dividing it into four unequally sized domains is cytoplasmic (Fig. 2d–f, i). Clockwise from the top left corner of the molecule depicted in Fig. 2a, the relative sizes of these domains are 27, 26, 25 and 22%.

Sections through the 3D density map reveal the internal structure of the voltage-sensitive sodium channel (Fig. 3b–g). Extracellular, transmembrane and cytoplasmic regions tentatively assigned by their surface structure (Fig. 2), and relative volumes (Fig. 3a) are

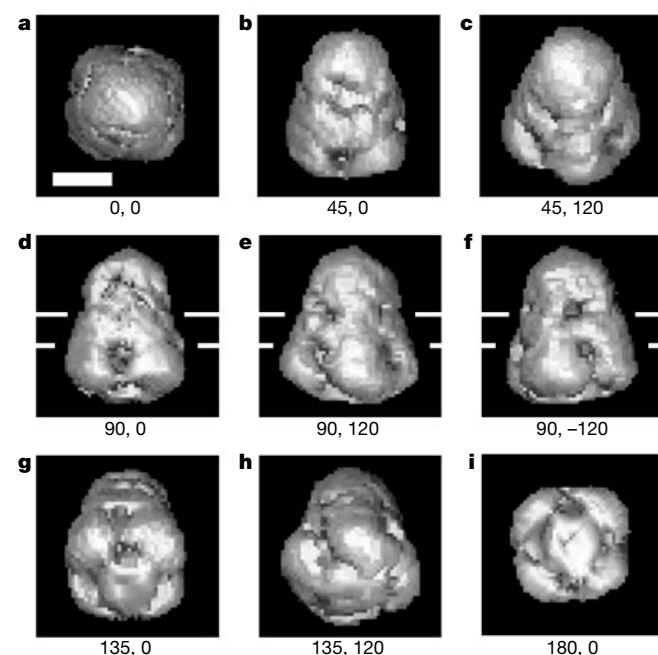


Figure 2 Surface representation of the sodium channel protein. **a**, Top view. **b**, View at an oblique angle, related to **a** by a +45° rotation about the horizontal axis. **c**, View at an oblique angle, related to **b** by a +120° rotation about the vertical axis. **d–f**, Three side-views rotated by 120° intervals around the vertical axis. **g**, View at an oblique angle, related to **d** by a +45° rotation about the horizontal axis. **h**, View at an oblique angle, related to **g** by +120° rotation about the vertical axis. **i**, Bottom view. The M_r enclosed by the isosurface is 320K, corresponding to 150% of the sodium channel protein. Euler angles (β , γ) are indicated; the α -angle is 0. White bars delineate the lipid bilayer region. Scale bar, 50 Å.

also distinct in the longitudinal sections shown in Fig. 3b–d. The cytoplasmic region houses one central and four peripheral cavities (Fig. 3b–e). The latter have a height and width of roughly 15 Å, and merge with the large cytosolic orifices distinguished in the surface reconstruction (Fig. 2). Four small orifices in the half-spherical extracellular structure (Figs 1–3) perforate the protein shell close to the putative transmembrane region and connect the large internal cavity with the exterior.

The large internal cavities on both the cytoplasmic and extracellular sides connect to four narrow, peripheral, low-density regions in the transmembrane domain (Fig. 3e–g). A rotation of their positions around the pseudo-four-fold axis, about 45° in the upper cross-section (Fig. 3g) relative to the lower cross-section (Fig. 3e), indicates that the low-density feature is twisted. Moreover, each low-density region exhibits a constriction at the level of the arrows in Fig. 3b–d. The distinct elongated central density has a length of 40 Å and a diameter of 35 Å, which is only slightly smaller than the dimensions of the KcsA pore-forming structure¹⁴.

We used antibody labelling of the C19 fragment¹³ of the C terminus (residues 1,802–1,820 of the protein sequence)¹⁵ to test the tentative domain assignments. Sodium channel/antibody complexes were separated from excess antibodies by gel-permeation chromatography and examined by negative-stain electron microscopy. The Y-shaped antibodies were often visible on the images, but their conformation was variable (Fig. 4a, b). To identify their binding site, similar end-on and side-view projections were aligned and averaged (Fig. 4c, e), using corresponding average projections of negatively stained sodium channels alone as references (Fig. 4d, f).

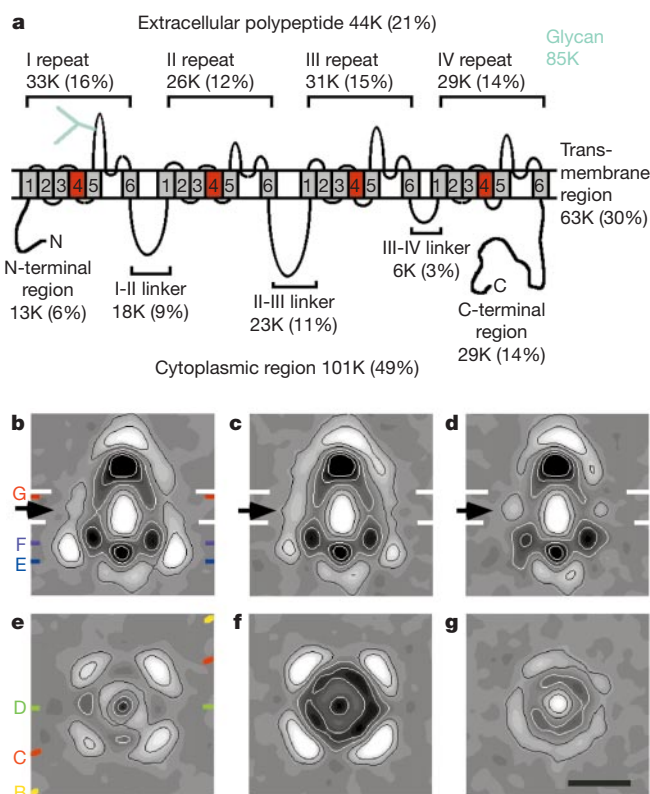


Figure 3 Domains and internal cavities of the sodium channel protein and the sequence-based structure prediction. **a**, Predicted repeats, loops and terminal extensions^{5,15}. The S4 segments are colored in red. **b**, Axial section along the diagonal indicated by the yellow line (B) in **e**. **c**, Axial section along the red line (C) in **e**. **d**, Axial section along the green line (D) in **e**. **e–g**, Perpendicular sections at the positions marked in **b** by lines coloured blue (E), purple (F) and red (G). High densities are shown in bright shades. White bars delineate the lipid bilayer. Black arrows indicate the constriction between the cytoplasmic (lower) and extracellular (upper) cavities. Scale bar, 50 Å.

Both average projections from the sodium channel/antibody complexes showed densities that could be related to the Fab arms of the antibodies: at the corner of the end-on view (Fig. 4c) and at the edge of the wide end of the side-view (Fig. 4e). The individual Fab domains were blurred, reflecting the flexible nature of the C terminus of the channel. As the C terminus is located on the cytoplasmic side of the membrane¹¹, the wide, square-shaped end of the sodium channel is indeed located in the cytoplasm.

Whereas voltage-sensitive potassium channels are homo-

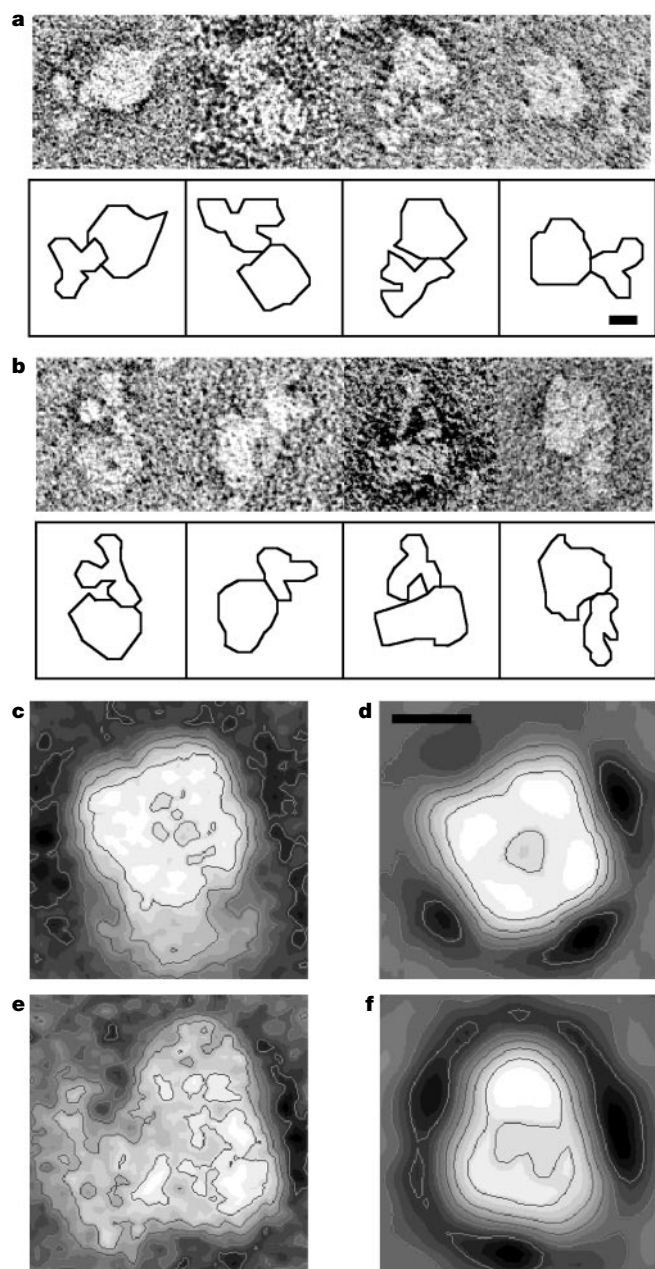


Figure 4 Electron microscopy of negatively stained sodium channel protein, anti-C-terminal antibody complexes and their average projections. **a**, End-on views. **b**, Side views. Contours are shown below each row to delineate the antibody and the sodium channel protein. **c**, **e**, Class averages of the sodium channel protein–antibody complexes; the end-on projection (**c**) shown with contours is the average of 142 projections, and the side view (**e**) is the average of 76 projections. **d**, **f**, Class averages of the sodium channel without antibody; the end-on (**d**) and side-view (**f**) projections are averages containing 196 and 132 projections, respectively. Protein is shown in bright shades. Scale bar, 50 Å.

tetramers²³, sodium and calcium channels are large monomeric proteins that include four homologous repeats^{15,16}. The high sequence homology^{15–17} of the transmembrane segments and S5–S6 linkers suggests that these channels share a similar structure. Comparison of the partial structure of KcsA¹⁴ and the lower-resolution 3D density map of the sodium channel presented here indicates that the prominent elongated central density visible in longitudinal sections of the latter (Fig. 3b–d) might correspond to the KcsA pore-forming structure. This is supported by the sequence homology between the KcsA protein and the S5–S6 linker and the S5 and S6 segments of the sodium channel¹⁴. However, neither the putative channel in the central density nor the features of the central cavity in the transmembrane region¹⁴ is revealed at 19 Å resolution. The low density at the four-fold symmetric axis of the negatively stained sodium channel¹³ (Fig. 4) is compatible with our current structure, although uranyl salts tend to fill hydrophilic cavities such as those revealed by cryo-electron microscopy. By cryo-electron microscopy, the protein moiety is delineated against the buffer containing 20% glycerol and 2 M MgCl₂, yielding a density of 1.185 g ml⁻¹. In contrast, for negatively stained samples flattening, uneven staining and reduced resolution should be considered; these factors explain minor differences between the projections of samples prepared differently.

The cavities and orifices of the sodium channel are unique and may be involved in voltage sensing and gating. In contrast, the acetylcholine receptor exhibits a single funnel-shaped ion inlet on the extracellular side²⁴. Cysteine scanning mutagenesis and accessibility tests using hydrophilic cysteine-modifying reagents have shown that only 2–5 residues of the hypothetical transmembrane segment IVS4 (Fig. 3a) are inaccessible²⁵. This agrees well with the internal hydrophilic environment created by the cavities. The conservation of hydrophilic residues in the putative transmembrane segments, for example, 36 positively and 22 negatively charged residues, is also compatible with the porous transmembrane structure. Because solubilized sodium channels were not exposed to a membrane potential, they were probably in their closed state on imaging. The short inaccessible hydrophobic region in IVS4 shifts markedly depending on the membrane potential²⁵. Conformational changes of the S5–S6 linker have also been observed during outer mouth gating²⁶. These observations, together with the presence of a single ion-selective filter revealed by mutagenesis⁹ and four voltage sensors^{8,15}, fosters the hypothesis that the ion pathway is through the elongated central density thought to include S6 (refs 1, 6), and that the pathway is opened by the central mass pushing out into the four surrounding cavities during voltage gating.

The technique that we have used, imaging in the helium-cooled cryo-electron microscope¹⁸ combined with the single particle analysis¹⁹, has provided a low-resolution 3D density map of the sodium channel protein that reveals complex internal cavities and pores. Further improvement of the resolution and structural analysis of sodium channels in the open state are now required to gain more detailed information on the structure of the complex cavity system and the mechanisms of ion transport and gating. □

Methods

Purification of the sodium channels

The voltage-sensitive sodium channel from the electric organ of *Electrophorus electricus* was purified by immunoaffinity chromatography and gel filtration as described¹³ except that the affinity chromatography gel contained 10 mg ml⁻¹ of purified antibody 1A. Elution from the antibody column¹³ was achieved by a step gradient containing 3.5 M MgCl₂. The subsequent gel permeation chromatography step was carried out as described¹³, but with buffer containing 2 M MgCl₂. For cryo-electron microscopy, the gel-permeation chromatography fraction of the sodium channels was either used directly or concentrated 2–3-fold by centrifugation in an Amicon centricon 100.

Cryo-electron microscopy

A 3-μl droplet of the sodium channel solution was applied to a holey carbon grid¹⁸. The grid was partially blotted with filter paper to remove excess solution, immediately plunged

into liquid ethane chilled by liquid nitrogen, transferred into the JEOL JEM3000SFF cryo-electron microscope¹⁸ and kept at 4.2 K. Low-dose (< 20 electrons per Å²) images were recorded on Kodak SO163 film at a nominal magnification of ×36,000 and with a 300-kV acceleration voltage. Applied underfocus values were 3.7 to 7.6 μm. Thirty-two images were digitized with a Scitex Leafscan 45 scanner²⁷ at a pixel size of 2.83 Å at the specimen level. The discernible projection images of the molecule were interactively extracted into 80 × 80 pixel subframes, and analysed after subtracting the uneven background.

Image analysis

Analysis of images was performed in three major refinement cycles using the Imagic V program system¹⁹. In the first cycle the selected 11,991 images were first aligned rotationally and translationally, and classified by the reference-free method, and subsequently grouped into 500 clusters by multivariate statistical analysis^{28,29}. The resulting averages were used as new references, and the cycle was repeated ten times. In the second cycle, original images were corrected for the contrast transfer function (CTF) of the electron microscope using the parameters Cs 1.6 mm, acceleration voltage 300 kV, Cc 2.2 mm, and aligned with the averages. This further refinement cycle, from the alignment to the generation of references, was repeated eight times. Before the third cycle, orientational Euler angles of the class averages were initially determined from the sinogram²² of cross-correlation functions assuming a C4 symmetry to obtain a preliminary 3D reconstruction by exact filtered back-projection³⁰. For the third cycle, aimed at refining the Euler angles, no symmetry was imposed.

We calculated 189 back-projections in evenly spaced directions from the 3D reconstruction. Each of the 11,991 CTF-corrected images was aligned against all 189 projections and subclustered, providing improved cluster averages, and a new 3D map was generated and projected as above. Again, this cycle was repeated until convergence. The final reconstruction included 11,543 images, 96.3% of the selected images. The isosurface threshold level was chosen to yield an essentially closed surface comprising a total volume of 4×10^5 Å³. To assess the resolution, independent reconstructions were calculated from two half-sets of the data and assessed by the Fourier shell correlation function³⁰.

Formation of the channel-antibody complex

The antibody 2B against the carboxy-terminal C19 fragment was prepared as described¹³. Formation of the sodium channel/antibody complex and removal of excessive antibody by gel-permeation chromatography were also carried out as described¹³, except that the antibody-sodium channel incubation and gel-permeation buffers were both supplemented with 0.1 M MgCl₂ and the incubation with antibody was made for 30 min at 4 °C. Negatively stained samples were recorded, digitized and processed as described¹³.

Received 4 September; accepted 22 December 2000.

- Catterall, W. A. From ionic currents to molecular mechanisms: the structure and function of voltage-gated sodium channels. *Neuron* **26**, 13–25 (2000).
- McClatchey, A. I. *et al.* Temperature-sensitive mutations in the III–IV cytoplasmic loop region of the skeletal muscle sodium channel gene in paramyotonia congenita. *Cell* **68**, 769–774 (1992).
- Ptacek, L. J. *et al.* Identification of a mutation in the gene causing hyperkalemic periodic paralysis. *Cell* **67**, 1021–1027 (1991).
- Deschenes, I. *et al.* Electrophysiological characterization of SCN5A mutations causing long QT (E1784K) and Brugada (R1512W and R1432G) syndromes. *Cardiovasc Res.* **46**, 55–65 (2000).
- James, W. M. & Agnew, W. S. Multiple oligosaccharide chains in the voltage-sensitive Na⁺ channel from *Electrophorus electricus*: evidence for α-2,8-linked polysialic acid. *Biochem. Biophys. Res. Commun.* **148**, 817–826 (1987).
- Ragsdale, D. S., McPhee, J. C., Scheuer, T. & Catterall, W. A. Molecular determinants of state-dependent block of Na⁺ channels by local anesthetics. *Science* **265**, 1724–1728 (1994).
- Kandel, E. R. & Schwartz, J. H. Molecular biology of learning: modulation of transmitter release. *Science* **218**, 433–443 (1982).
- Stühmer, W. *et al.* Structural parts involved in activation and inactivation of the sodium channel. *Nature* **339**, 597–603 (1989).
- Heinemann, S. H., Terlau, H., Stühmer, W., Imoto, K. & Numa, S. Calcium channel characteristics conferred on the sodium channel by single mutations. *Nature* **356**, 441–443 (1992).
- Vassilev, P. M., Scheuer, T. & Catterall, W. A. Identification of an intracellular peptide segment involved in sodium channel inactivation. *Science* **241**, 1658–1661 (1988).
- Gordon, R. D., Fieles, W. E., Schotland, D. L., Hogue-Angeletti, R. & Barchi, R. L. Topographical localization of the C-terminal region of the voltage-dependent sodium channel from *Electrophorus electricus* using antibodies raised against a synthetic peptide. *Proc. Natl Acad. Sci. USA* **84**, 308–312 (1987).
- Rohli, C. A. *et al.* Solution structure of the sodium channel inactivation gate. *Biochemistry* **38**, 855–861 (1999).
- Sato, C., Sato, M., Iwasaki, A., Doi, T. & Engel, A. The sodium channel has four domains surrounding a central pore. *J. Struct. Biol.* **121**, 314–325 (1998).
- Doyle, D. A. *et al.* The structure of the potassium channel: molecular basis of K⁺ conduction and selectivity. *Science* **280**, 69–77 (1998).
- Noda, M. *et al.* Primary structure of *Electrophorus electricus* sodium channel deduced from cDNA sequence. *Nature* **312**, 121–127 (1984).
- Tanabe, T. *et al.* Primary structure of the receptor for calcium channel blockers from skeletal muscle. *Nature* **328**, 313–318 (1987).
- Tempel, B. L., Papazian, D. M., Schwarz, T. L., Jan, Y. N. & Jan, L. Y. Sequence of a probable potassium channel component encoded at Shaker locus of *Drosophila*. *Science* **237**, 770–775 (1987).
- Fujiyoshi, Y. The structural study of membrane proteins by electron crystallography. *Adv. Biophys.* **35**, 25–80 (1998).

- van Heel, M., Harauz, G., Orlova, E. V., Schmidt, R. & Schatz, M. A new generation of the IMAGIC image processing system. *J. Struct. Biol.* **116**, 17–24 (1996).
- Penczek, P., Radermacher, M. & Frank, J. Three-dimensional reconstruction of single particles embedded in ice. *Ultramicroscopy* **40**, 33–53 (1992).
- van Heel, M. Classification of very large electron microscopical image data sets. *Optik* **82**, 114–126 (1989).
- van Heel, M. Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. *Ultramicroscopy* **38**, 241–251 (1987).
- MacKinnon, R. Determination of the subunit stoichiometry of a voltage-activated potassium channel. *Nature* **350**, 232–235 (1991).
- Miyazawa, A., Fujiyoshi, Y., Stowell, M. & Unwin, N. Nicotinic acetylcholine receptor at 4.6 Å resolution: Transverse tunnels in the channel wall. *J. Mol. Biol.* **288**, 765–786 (1999).
- Yang, N., George, A. L. Jr & Horn, R. Molecular basis of charge movement in voltage-gated sodium channels. *Neuron* **16**, 113–122 (1996).
- Liu, Y., Jurman, M. E. & Yellen, G. Dynamic rearrangement of the outer mouth of a K⁺ channel during gating. *Neuron* **16**, 859–867 (1996).
- Mitsuoka, K., Murata, K., Kimura, Y., Namba, K. & Fujiyoshi, Y. Examination of the LeafScan 45, a line-illuminating micro-densitometer, for its use in electron crystallography. *Ultramicroscopy* **68**, 109–121 (1997).
- van Heel, M. Multivariate statistical classification of noisy images (randomly oriented biological macromolecules). *Ultramicroscopy* **13**, 165–183 (1984).
- Frank, J., Bretaudiere, J. P., Carazo, J. M., Verschoor, A. & Wagenknecht, T. Classification of images of biomolecular assemblies: a study of ribosomes and ribosomal subunits of *Escherichia coli*. *J. Microsc.* **150**, 99–115 (1988).
- Harauz, G. & van Heel, M. Exact filters for general geometry three dimensional reconstruction. *Optik* **73**, 146–156 (1986).

Acknowledgements

We thank S. Müller for her illuminating suggestions and help with the manuscript, and M. van Heel, M. Schatz and R. Schmidt for their helpful and constructive advice. T. Moriya promoted the present research. We also thank K. Imoto for discussions, A. Oshima for suggestions and S. Miyazaki for assistance. This work was supported by grants from the ETL, the Japan New Energy and Industrial Technology Development Organization (NEDO), and the Swiss National Science Foundation (to A.E.), and by the Maurice E. Müller Foundation, Switzerland.

Correspondence and requests for materials should be addressed to C.S. (e-mail: tsisato@etl.go.jp).

RGS2 regulates signal transduction in olfactory neurons by attenuating activation of adenylyl cyclase III

Srikumar Sinnarajah*, Carmen W. Dessauer†, Deepa Srikumar*, Jun Chen†, John Yuen*, Solomon Yilma‡, John C. Dennis‡, Edward E. Morrison‡, Vitaly Vodanoy‡ & John H. Kehrl*

* B cell Molecular Biology Section, Laboratory of Immunoregulation, NIAID, NIH, Bethesda, Maryland 20892, USA

† Department of Integrative Biology and Pharmacology, University of Texas-Houston Medical School, Houston, Texas 77225, USA

‡ Department of Anatomy, Physiology, and Pharmacology, Auburn University, Auburn, Alabama 36849, USA

The heterotrimeric G-protein G_s couples cell-surface receptors to the activation of adenylyl cyclases and cyclic AMP production (reviewed in refs 1, 2). RGS proteins, which act as GTPase-activating proteins (GAPs) for the G-protein α-subunits α_i and α_q, lack such activity for α_s (refs 3–6). But several RGS proteins inhibit cAMP production by G_s-linked receptors^{7,8}. Here we report that RGS2 reduces cAMP production by odorant-stimulated olfactory epithelium membranes, in which the α_s family member α_{olf} links odorant receptors to adenylyl cyclase activation^{9,10}. Unexpectedly, RGS2 reduces odorant-elicited cAMP production, not by acting on α_{olf} but by inhibiting the activity of adenylyl cyclase type III, the predominant adenylyl cyclase isoform in olfactory neurons. Furthermore, whole-cell voltage clamp recordings of odorant-stimulated olfactory

neurons indicate that endogenous RGS2 negatively regulates odorant-evoked intracellular signalling. These results reveal a mechanism for controlling the activities of adenylyl cyclases, which probably contributes to the ability of olfactory neurons to discriminate odours.

In the nasal cavity, volatile chemicals (odorants) bind seven transmembrane receptors (olfactory receptors) present in the olfactory epithelium, stimulating α_{olf} nucleotide exchange and the dissociation of α_{olf} from the $\beta\gamma$ subunits (refs 9, 10). GTP· α_{olf} binds and activates adenylyl cyclase type III thereby increasing intracellular cAMP levels, which triggers the opening of cyclic-nucleotide-gated cation channels, membrane depolarization, and the generation of action potentials^{11,12}. GTP hydrolysis halts α_{olf} signalling and facilitates the re-assembly of the heterotrimeric G_{olf}. As several RGS proteins inhibit signalling through α_s -coupled receptors^{7,8}, here we examined their effect on α_{olf} signalling.

We measured cAMP production by olfactory epithelium membranes stimulated with odorants in the presence of GTP γ S, a non-hydrolyzable form of GTP. Odorants triggered a 15-fold increase in cAMP, whereas similarly prepared respiratory epithelium membranes did not respond (data not shown). Adding recombinant RGS1, RGS2 or RGS3 decreased odorant-induced cAMP production, whereas the addition of two other RGS proteins, RGS4 and RGS5, did not (Fig. 1a). Among the three inhibitory RGS proteins, RGS2 was the most effective: 100 nM RGS2 reduced cAMP production by 50% (Fig. 1b).

As we used GTP γ S in these experiments, it is unlikely that RGS2 acted as a GAP for α_{olf} . Also, RGS2 similarly reduced GTP γ S· α_s -stimulated cAMP production by olfactory membranes (see below), eliminating the possibility that RGS2 selectively inhibited α_{olf} -triggered production of cAMP. To test whether RGS2 directly blocked α_s from activating adenylyl cyclases, we pre-incubated either GTP γ S· α_s or the olfactory membranes with RGS2 before measuring cAMP production. Unexpectedly, pre-incubation of

the membranes inhibited cAMP production much more than did pre-incubation of GTP γ S· α_s (Fig. 2a). RGS2 seemed to reduce adenylyl cyclase activation independently of any effect on α_s or α_{olf} . Consistent with this possibility, RGS2 inhibited the production of cAMP by olfactory membranes triggered by forskolin (Fig. 2b), an adenylyl cyclase activator that binds at a site distinct from the binding site of GTP· α_s (ref. 2). RGS2 also reduced cAMP production by forskolin- or GTP γ S· α_s -treated Sf9 membranes engineered to express adenylyl cyclase type III (Fig. 2c).

Using a similar approach, we determined whether other adenylyl cyclase isoforms shared this sensitivity to RGS2. RGS2 inhibited cAMP production by GTP γ S· α_s - or forskolin-stimulated Sf9 membranes expressing type V or VI adenylyl cyclases, but not by those expressing type I or II (Fig. 3a, b). We did not test whether RGS2 inhibits adenylyl cyclase types IV, VII, VIII and IX. Next, using the purified cytoplasmic domains of type V adenylyl cyclase¹³, we tested whether RGS2 directly inhibited adenylyl cyclase activity. The addition of purified recombinant RGS2 decreased cAMP production by GTP γ S· α_s - or forskolin-stimulated soluble type V adenylyl

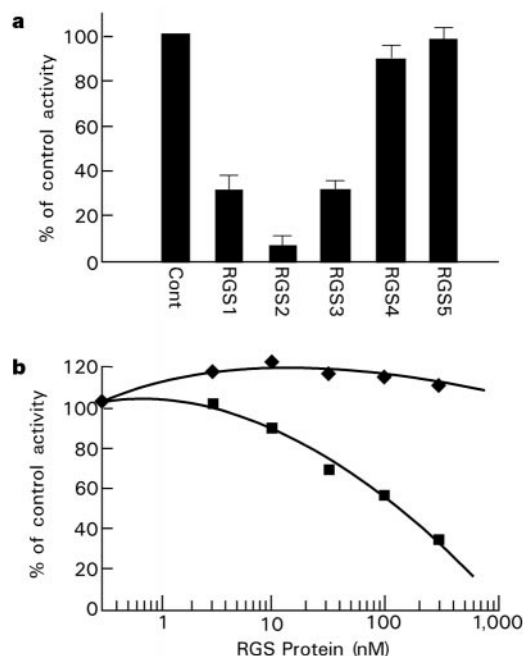


Figure 1 Effects of RGS proteins on adenylyl cyclase activity. **a**, Olfactory membranes were pre-incubated with the indicated RGS proteins (1 μ M) and activated with the odorant mixture and GTP γ S. cAMP levels from three experiments are expressed as a percentage of the control value (no RGS protein). **b**, Olfactory epithelium membranes were pre-incubated with various concentrations of RGS2 (filled squares) or RGS4 (filled diamonds), and activated with the odorant mixture plus GTP γ S. cAMP levels are expressed as a percentage of the control value.

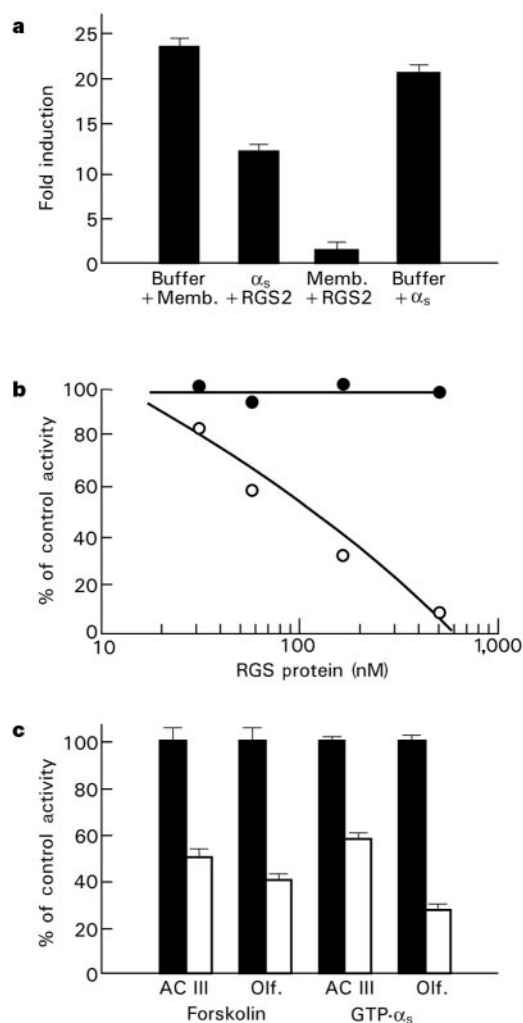


Figure 2 RGS2 inhibits odorant- and forskolin-induced cAMP production. **a**, GTP γ S· α_s or olfactory membranes were pre-incubated for 30 min with RGS2 (250 nM) or buffer and then combined with membranes or GTP γ S· α_s , respectively. After 5 min cAMP production was measured. **b**, Forskolin (10 μ M) stimulated olfactory membranes previously pre-incubated with either RGS2 (filled squares) or RGS4 (filled circles). cAMP levels are expressed as a percentage of the control value (no RGS proteins). **c**, Olfactory membranes (Olf.) or Sf9 membranes expressing type III adenylyl cyclase (AC III) were pre-incubated in the absence (black bar) or presence of 500 nM RGS2 (white bar), and assayed for cAMP production in the presence of GTP γ S· α_s or forskolin.

cyclase, whereas the addition of recombinant RGS4 did not (Fig. 3c). Thus, RGS2 directly inhibits the activities of adenylyl cyclase type V, and probably types III and VI, but types I and II are insensitive to RGS2.

To verify that RGS2 inhibits α_s -induced adenylyl cyclase type III activation *in vivo*, we transfected HEK 293 cells with expression vectors for adenylyl cyclase type III and a GTPase-deficient form of α_s , α_s Q227L, in the presence or absence of RGS2. The cellular expression of α_s Q227L causes activation of adenylyl cyclases and cAMP accumulation¹⁴. Co-expression of RGS2 inhibited the augmented cAMP production that we observed with the addition of adenylyl cyclase III (Fig. 3d); however, RGS2 had little effect on the α_s Q227L-induced cAMP accumulation that occurred without adenylyl cyclase III expression. HEK 293 cells do not express adenylyl cyclase III, but presumably express adenylyl cyclase isoforms that are largely insensitive to RGS2.

Because of the pronounced effect of RGS2 on odorant-induced cAMP production, we examined olfactory epithelium for RGS2 expression. We detected RGS2 in olfactory epithelium cell lysates by immunoblot (data not shown) and in most olfactory neuron cell bodies by immunocytochemistry. An antibody specific for olfactory marker protein confirmed the presence of olfactory epithelium, whereas a pre-immune serum showed little reactivity (Fig. 4a–c). Previous immunocytochemical studies detected adenylyl cyclase type III primarily in the apical ciliary layer¹⁵, where a proportion of RGS2 also localized.

Next, we performed whole-cell voltage clamp recordings of olfactory neurons in the presence or absence of the RGS2-specific antibodies. In their absence, odorants elicited characteristic membrane depolarization and repolarization¹⁶, which depended on the applied voltage, whereas cells from the respiratory epithelium did not respond (Fig. 4d, e). When we filled the patch electrode with a

solution containing RGS2 antibodies, the odorant-evoked currents changed markedly. Five minutes after beginning the perfusion, we noted a progressive increase in the amplitude of the odorant triggered inward currents (Fig. 4f). The enhancement of odorant-induced inward currents by RGS2 antibodies did not depend on the opening of sodium channels, as the sodium channel blocker tetrodotoxin had no appreciable effect on them (Fig. 4g). Pre-incubation with recombinant RGS2 blocked their enhancement of odorant-induced inward currents (Fig. 4h), indicating that the RGS2 antibodies reacted with endogenous RGS2. In addition, control antibodies had no effect on the odorant-induced inward currents.

Normally, odorants raise intracellular cAMP levels, which opens cyclic nucleotide-gated cation channels resulting in an influx of Ca^{2+} and the opening of Ca^{2+} -activated Cl^- channels^{17,18}. The diffusion of RGS2 antibodies into olfactory neurons should progressively neutralize RGS2, thereby removing its tempering effect on α_{olf} -stimulated adenylyl cyclase III activity, leading to elevated cAMP levels, opening of additional cyclic-nucleotide-gated channels, and increased inward currents as we observed. An alternative explanation for the increased odorant-elicited inward currents observed with the RGS2 antibodies is the neutralization of RGS2 α_{olf} GAP activity; however, our biochemical studies do not support this possibility. In addition, the effects of the RGS2 antibody cannot be readily explained by neutralization of RGS2 α_i GAP activity, as this neutralization would be predicted to decrease the inward current.

The sensitivity of adenylyl cyclase III to RGS2 suggests that the constitutive level of RGS2 in olfactory neurons may set a threshold for α_{olf} -mediated signalling. Furthermore, as the two main second messengers generated by odorants, cAMP and Ca^{2+} , both elevate RGS2 messenger RNA expression in other cell types^{19,20}, RGS2 might

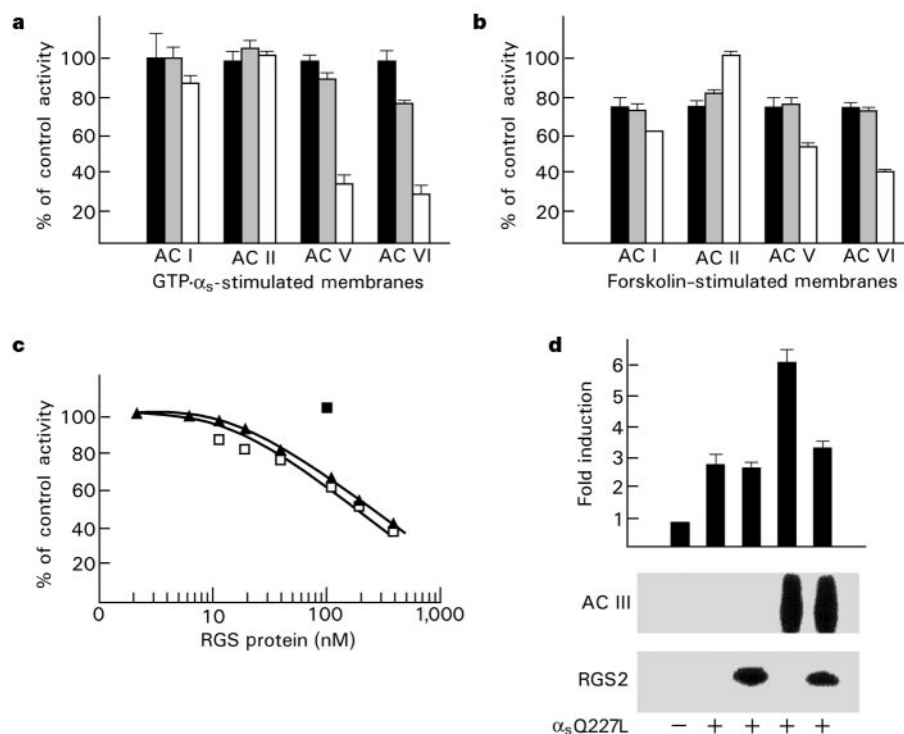


Figure 3 RGS2 inhibits the activity of specific adenylyl cyclases. **a, b**, cAMP production by Sf9 membranes expressing the indicated adenylyl cyclase incubated with buffer (black bars), 150 nM (dotted bars) or 500 nM (white bars) RGS2 and stimulated with GTP- γ - α_s (**a**) or forskolin (**b**). Data are expressed as a percentage of the control value. **c**, Effect of RGS2 on cAMP production by the C₁ and C₂ domains of adenylyl cyclase type V stimulated

with GTP- γ - α_s (open squares) or forskolin (filled triangles). Similar assays performed with a single concentration of RGS4 (filled square). **d**, Production of cAMP by HEK 293 cells transfected as indicated with adenylyl cyclase III, α_s Q227L and RGS2. Fold inductions are the mean $\pm$ two s.d. of three experiments. Adenylyl cyclase III and RGS2 (HA) levels were assayed by immunoblot.

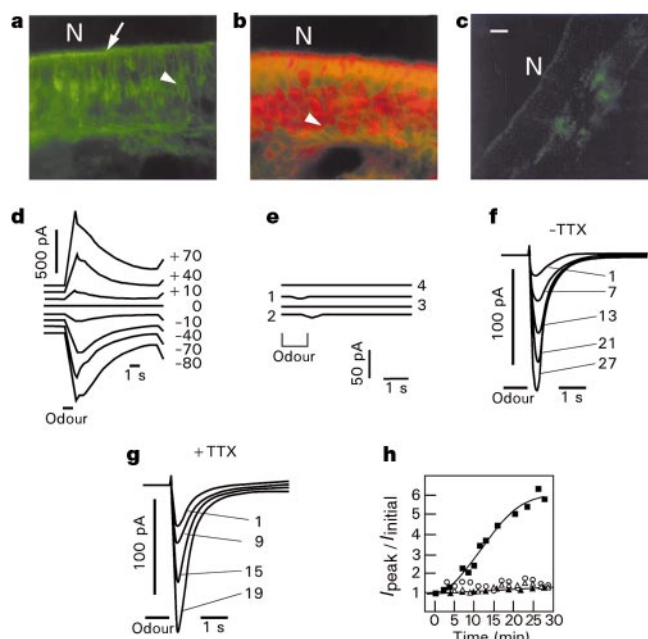


Figure 4 RGS2 attenuates odorant signalling in olfactory neurons. **a–c**, Immunocytochemistry with RGS2-specific antibodies (**a**), or antibodies directed against olfactory marker protein (**b**), or a control (**c**). Arrow in **a** indicates RGS2 expression in the apical ciliary layer. Arrowheads in **a** and **b** delineate individual olfactory neurons. N, nasal cavity. Scale bar, 20 μm . **d**, Current traces from whole-cell voltage recordings of olfactory neurons with holding potentials from -80 to $+70$ mV. The stimulus was a 1-s pulse of odorant mixture. Downward direction indicates inward current. **e**, Current traces from whole-cell voltage clamp recordings of respiratory epithelium neurons with holding potentials at 0 mV (1) and 90 mV (2), and for patch electrodes contacting a buffer solution with potentials at 0 mV (3) and 60 mV (4). **f**, Odorant-induced currents recorded with

patch electrodes filled with RGS2 antibodies ($4 \mu\text{g ml}^{-1}$). Odorant puffs at 1, 7, 13, 21 and 27 min after initiating the perfusion as indicated. The cell was clamped at -10 mV.

g, Similar experiment as **f** except that 100 nM tetrodotoxin was added to the buffer solution 5 min before odorant. Odorant puffs at 1, 9, 15 and 19 min after initiating the perfusion. **h**, Plot of normalized, peak negative current ($I_{\text{peak}}/I_{\text{initial}}$) evoked by odorant versus perfusion time with RGS2 antibody (filled squares), $I_{\text{initial}} = -22.0$ pA; a mixture of RGS2 antibody and RGS2 protein (filled triangles), $I_{\text{initial}} = -53.2$ pA; buffer (open triangles), $I_{\text{initial}} = -14.3$ pA; or anti-olfactory marker protein antibody (open circles), $I_{\text{initial}} = -94.7$ pA.

mediate long-term adaptation to odorants, although perhaps not rapid attenuation, which has been attributed to the phosphorylation and inhibition of adenylyl cyclase by CaM kinase II (ref. 21). In addition, because cardiac myocytes express abundantly two RGS2-sensitive adenylyl cyclase isoforms (types V & VI)¹, cardiac RGS2 expression may temper the positive inotropic effect exerted by β -agonists. □

Methods

Plasmids and antibodies

Expression vectors for the RGS proteins and adenylyl cyclases have been described¹⁴. The αsQ227L expression vector was provided by S. Gutkind (NIH). The anti-HA (haemagglutinin A) and anti-ACIII antibodies were purchased from Covance and Santa Cruz, respectively. The anti-RGS2 antiserum was made in rabbits against a carboxy-terminal peptide (PQITTEPHAT) coupled to keyhole limpet haemocyanin followed by affinity purification using the immunizing peptide. The anti-olfactory marker protein antibody was a gift from F. Margolis (Univ. Maryland School of Medicine).

Transfection of HEK 293 cells

HEK 293 cells were transfected using calcium phosphate with constructs expressing αsQ227L (0.5 μg per plate), RGS2 (1 μg per plate) and adenylyl cyclase III (0.2 μg per plate). The cells were collected 24 h later, and lysates were tested for the production of cAMP (see below). We monitored protein expression by immunoblot.

Olfactory epithelium membrane preparation

We prepared membranes by a modification of published procedures²². Dissected rat olfactory epithelium tissues, in EDTA/Ringer's solution containing 2 mM HEPES, 112 mM NaCl, 3.4 mM KCl, 2.4 mM NaHCO_3 and 2 mM EDTA, pH 7.4, were sonicated for 2 min (Microson tip sonicator at 10 Watts, 23 kHz, at 4°C under nitrogen gas). The de-ciliated epithelia were sedimented, and the supernatant centrifuged at 3,000g for 5 min and at 146,000g for 30 min, both spins at 4°C . Purified cilia membrane protein (1 μg) was incubated for 5 min at 37°C in 50 mM Tris-HCl, 5 mM MgCl_2 , 20 mM creatine phosphate, 50 units ml^{-1} of creatine kinase and 2 mg ml^{-1} of bovine serum albumin. RGS proteins were pre-incubated with the olfactory epithelium membranes for 30 min at 22°C . The

samples were activated with an odorant mixture (25 μM of equal parts ethyl butyrate, eugenol, and (+) and (–) carvones), $\alpha\text{s} \cdot \text{GTP}\gamma\text{S}$, or forskolin for 5 min at 37°C (all stimulants included 2 mM ATP). The reactions were stopped by boiling, and centrifuged at 13,800g for 5 min. cAMP levels were measured by radioimmunoassay (cAMP RIA). Alternatively, boiled lysates from transfected cells were used for the cAMP RIA. Thirty minutes before lysis, the cells were exposed to 1 mM 3-isobutyl-1-methylxanthine in DMEM plus 1% fetal calf serum to inhibit endogenous phosphodiesterases.

Purification of recombinant RGS proteins

Polymerase chain reaction fragments prepared from RGS plasmids were inserted in-frame with hexa-histidine tag in pET15b. IPTG-induced RGS proteins were purified under non-denaturing conditions. The culture of Sf9 (*Spodoptera frugiperda*) cells, and the production, cloning and amplification of recombinant baculoviruses were performed as described²³. The membranes were collected, washed and re-suspended in 20 mM sodium HEPES, pH 8.0, 2 mM dithiothreitol and 200 mM sucrose.

Purification and assay of C1 and C2 domains

Adenylyl cyclase domains were expressed with an amino-terminal (VC2) or C-terminal (VC1(670)) hexa-histidine tag and purified¹³. Synthesis of cAMP by the C1 and C2 domains was measured at 30°C for 8–10 min with 1 mM $[\gamma\text{-}^{32}\text{P}]\text{ATP}$, 5 mM MgCl_2 and 50 μM forskolin or 0.4 μM $\text{GTP}\gamma\text{S} \cdot \alpha\text{s}$. After quenching the products were separated by sequential chromatography²⁴. Activities are expressed per mg of the limiting adenylyl cyclase domain in the assay (C_1). The other domain (C_2) was present in excess (0.5 μM) to drive the interaction between the two proteins.

Immunocytochemistry

Rat ethmoturbinate were fixed with 4% paraformaldehyde, decalcified in 0.25 M EDTA, passed through sucrose solutions of 10%, 20% and 30% placed in plastic moulds that contained Tissue-Tek OCT embedding medium (Sakura) and frozen. Sections (10 μm) were cut, placed on poly-L-coated (Sigma) glass slides, and air dried for 1 h. After washes with distilled H_2O and PBS, the slides were blocked with 10% normal serum and 5% BSA. The primary antibody diluted in the blocking solution was applied overnight, the slides were washed, and Oregon-green-conjugated goat anti-rabbit IgG (1:100; Molecular Probes) or TRITC-conjugated anti-goat IgG (Boehringer Mannheim) was applied for 1 h. The slides were observed through an epi-fluorescence-equipped Nikon Optiphot 2

microscope. Images were captured with a RT-Slider Spot digital camera (Diagnostic Instruments).

Whole-cell voltage clamp recordings

Rat olfactory epithelia were used for whole-cell voltage clamp recordings²⁵. Fragments of septal olfactory mucosa were placed in the perfusion chamber such that the basal portions were immersed in physiologic buffer (136.9 mM NaCl, 5.3 mM KCl, 4.2 mM NaHCO₃, 0.4 mM KH₂PO₄, 3.4 mM Na₂HPO₄, 5.6 mM D-glucose, pH 7.4), whereas the upper epithelial surfaces with olfactory cilia were exposed to the air. The patch electrode (resistance of 8–16 MΩ) was filled with a solution containing 110 mM KCl, 4 mM NaCl, 2 mM NaHCO₃, 1 mM MgCl₂, 0.1 mM CaCl₂ and 2 mM MOPS at pH 7.4 in the presence or absence of affinity purified RGS2 antibodies (4 μg ml⁻¹), RGS2 antibodies (4 μg ml⁻¹) plus recombinant RGS2 (4 μg ml⁻¹) or control antibodies. After stable contact with an olfactory neuron, 1-s puffs of odorant mixture containing 1.6 mM ethyl butyrate, eugenol and (+) and (-) carvones were applied. The whole-cell response over the time course of several minutes was recorded after being amplified by a voltage-clamp amplifier (Axiopatch 200B, Axon Inst.) and filtered at 2–5 kHz. After compensation the series resistance was always lower than 20 MΩ. In some experiments the epithelia were incubated for 5 min in buffer containing tetrodotoxin.

Received 8 August; accepted 4 December 2000.

1. Sunahara, R. K., Dessauer, C. W. & Gilman, A. G. Complexity and diversity of mammalian adenylyl cyclases. *Annu. Rev. Pharmacol. Toxicol.* **36**, 461–480 (1996).
2. Hurlley, J. H. Structure, mechanism, and regulation of mammalian adenylyl cyclase. *J. Biol. Chem.* **274**, 7599–7602 (1999).
3. Hepler, J. R. Emerging roles for RGS proteins in cell signaling. *Trends Pharmacol. Sci.* **20**, 376–382 (1999).
4. Kehrl, J. H. Heterotrimeric G protein signaling: roles in immune function and fine-tuning by RGS proteins. *Immunity* **8**, 1–10 (1998).
5. Sunahara, R. K., Tesmer, J. J., Gilman, A. G. & Sprang, S. R. Crystal structure of the adenylyl cyclase activator Gsα. *Science* **278**, 1943–1947 (1997).
6. Natochin, M. & Artemyev, N. O. A single mutation Asp229→Ser confers upon Gsα the ability to interact with regulators of G protein signaling. *Biochemistry* **37**, 13776–13780 (1998).
7. Chatterjee, T. K., Eapen, A. K. & Fisher, R. A. A truncated form of RGS3 negatively regulates G protein-coupled receptor stimulation of adenylyl cyclase and phosphoinositide phospholipase C. *J. Biol. Chem.* **272**, 15481–15487 (1997).
8. Tseng, C. C. & Zhang, X. Y. Role of regulator of G protein signaling in desensitization of the glucose-dependent insulinotropic peptide receptor. *Endocrinology* **139**, 4470–4475 (1998).
9. Schild, D. & Restrepo, D. Transduction mechanisms in vertebrate olfactory receptor cells. *Physiol. Rev.* **78**, 429–466 (1998).
10. Reed, R. R. Signaling pathways in odorant detection. *Neuron* **8**, 205–209 (1992).
11. Pace, U., Hanski, E., Salomon, Y. & Lancet, D. Odorant sensitive adenylyl cyclase may mediate olfactory reception. *Nature* **316**, 255–258 (1985).
12. Sklar, P. B., Anholt, R. R. & Synder, S. H. The odorant-sensitive adenylyl cyclase of olfactory receptor cells. Differential stimulation by distinct classes of odorants. *J. Biol. Chem.* **261**, 15538–15543 (1986).
13. Dessauer, C. W., Tesmer, J. J., Sprang, S. R. & Gilman, A. G. Identification of a Gα binding site on type V adenylyl cyclase. *J. Biol. Chem.* **273**, 25831–25839 (1998).
14. Landis, C. A. *et al.* GTPase inhibiting mutations activate the alpha chain of Gs and stimulate adenylyl cyclase in human pituitary tumours. *Nature* **340**, 692–696 (1989).
15. Bakalyar, H. A. & Reed, R. R. Identification of a specialized adenylyl cyclase that may mediate odorant detection. *Science* **250**, 1403–1406 (1990).
16. Firestein, S. & Werblin, F. Odor-induced membrane currents in vertebrate olfactory receptor neurons. *Science* **244**, 79–82 (1989).
17. Kleene, S. J. & Gesteland, R. C. Calcium-activated chloride conductance in frog olfactory cilia. *J. Neurosci.* **6**, 3624–3629 (1991).
18. Kurahashi, T. & Taw, K.-W. Co-existence of cationic and chloride components in odorant-induced current of vertebrate olfactory receptor cells. *Nature* **363**, 71–74 (1993).
19. Pepperl, D. J., Shah-Basu, S., VanLeeuwen, D., Granneman, J. G. & MacKenzie, R. G. Regulation of RGS mRNAs by cAMP in PC12 cells. *Biochem. Biophys. Res. Commun.* **243**, 52–55 (1998).
20. Beadling, C., Druey, K. M., Richter, G., Kehrl, J. H. & Smith, K. A. Regulators of G protein signaling exhibit distinct patterns of gene expression and target G protein specificity in human lymphocytes. *J. Immunol.* **162**, 2677–2682 (1999).
21. Wei, J. *et al.* Phosphorylation and inhibition of olfactory adenylyl cyclase by CaM kinase II in Neurons: a mechanism for attenuation of olfactory signals. *Neuron* **21**, 495–504 (1998).
22. Sinnarajah, S. *et al.* Inhibition and enhancement of odorant-induced cAMP accumulation in rat olfactory cilia by antibodies directed against Gαs/olf- and Gαi-protein subunits. *FEBS Lett.* **426**, 377–380 (1998).
23. Summers, M. D. & Smith, G. E. in *Texas Agricultural Experiment Station Bulletin* 1,555 (College Station, Texas, 1987).
24. Salomon, Y., Londos, C. & Rodbell, M. A highly sensitive adenylyl cyclase assay. *Anal. Biochem.* **58**, 541–548 (1974).
25. Firestein, S. & Werblin, F. in *Chemical Senses Vol. 1, Receptor Events and Transduction in Taste and Olfaction* (eds Brand, J. G. *et al.*) 449–467 (Marcel Dekker, New York, 1989).

Acknowledgements

We thank M. Rust for editorial assistance; K. Druey, H. Cho and A. Sheschonka for the recombinant RGS proteins; and A. Fauci for support. E.M. and V.V. were supported by grants from the Office of Special Technology, Federal Aviation Administration and Defense Advanced Research Project Agency. C.W.D. was supported by the American Heart Association and the National Institutes of Health.

Correspondence and requests for materials should be addressed to J.K. (e-mail: Jkehrl@niaid.nih.gov).

Recognition of haemagglutinins on virus-infected cells by NKp46 activates lysis by human NK cells

Ofer Mandelboim*, Niva Lieberman*, Marianna Lev†, Lada Paul†, Tal I. Arnon*, Yuri Bushkin‡, Daniel M. Davis§, Jack L. Strominger||, Jonathan W. Yewdell¶ & Angel Porgador†

* The Lautenberg Center for General and Tumor Immunology, The Hebrew University-Hadassah Medical School, Jerusalem 91120, Israel

† Department of Microbiology and Immunology, Faculty of Health Sciences, and the Cancer Research Center, Ben Gurion University of the Negev, Beer Sheva 84105, Israel

‡ Laboratory of Molecular Immunology, Public Health Research Institute, 455 First Avenue, New York, New York 10016, USA

§ Department of Biology, Imperial College of Science, Technology and Medicine, London SW7 2AZ, UK

|| Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 01238, USA

¶ Laboratory of Viral Diseases, National Institute of Allergy and Infectious diseases, National Institutes of Health, Bethesda, Maryland

Natural killer (NK) cells destroy virus-infected and tumour cells, apparently without the need for previous antigen stimulation¹. In part, target cells are recognized by their diminished expression of major histocompatibility complex (MHC) class I molecules, which normally interact with inhibitory receptors on the NK cell surface^{2–8}. NK cells also express triggering receptors that are specific for non-MHC ligands; but the nature of the ligands recognized on target cells is undefined^{9–14}. NKp46 is thought to be the main activating receptor for human NK cells^{9,15}. Here we show that a soluble NKp46–immunoglobulin fusion protein binds to both the haemagglutinin of influenza virus and the haemagglutinin–neuraminidase of parainfluenza virus. In a substantial subset of NK cells, recognition by NKp46 is required to lyse cells expressing the corresponding viral glycoproteins. The binding requires the sialylation of NKp46 oligosaccharides, which is consistent with the known sialic binding capacity of the viral glycoproteins. These findings indicate how NKp46-expressing NK cells may recognize target cells infected by influenza or parainfluenza without the decreased expression of target-cell MHC class I protein.

We studied the role of NKp46 in NK recognition by producing a fusion protein in which the extracellular domain of NKp46 is fused to the Fc portion of immunoglobulin (Ig). Previous reports suggested that NKp46, together with the NKp44 activating receptor, is involved in the lysis of MHC class-I-negative 721.221 cells^{12,14,15}. We observed little staining of 721.221 cells when cells were incubated with the NKp46–Ig fusion protein (Table 1). As NK cells can effectively lyse virus-infected cells¹, we tested whether infection with Sendai virus (SV; a mouse paramyxovirus) increased the binding of NKp46–Ig. Remarkably, a 10-fold increase in the staining by NKp46–Ig was observed (Table 1). This effect is specific for NKp46, as SV infection did not alter the binding of other NK receptor–Ig fusion proteins tested (NKAT-8–Ig (KIR2DS4–Ig), KIR-1–Ig (KIR2DL1–Ig), or CD16–Ig; data not shown).

To identify the putative NKp46 ligand, we immunized mice with SV-infected 721.221 cells, and screened spleen-derived B-cell hybridoma supernatants for increased staining of virus-infected cells relative to non-infected cells. The supernatants of one of the hybridoma clones tested (135.7) efficiently blocked the binding of NKp46–Ig to SV-infected cells (Table 1). Enzyme-linked immunosorbent assays (ELISAs) using SV as immunoabsorbent indicated that monoclonal antibody (mAb) 135.7 recognizes a viral gene

product (data not shown).
SDS–polyacrylamide gel electrophoresis (PAGE) analysis of 135.7-reactive proteins immunoprecipitated from detergent lysates of ¹²⁵I cell-surface-labelled SV-infected 721.221 cells revealed that 135.7 binds a protein with an apparent relative molecular mass of 70,000 (*M_r* 70K), similar to the reported mobility of the haemagglutinin–neuraminidase (HN) glycoprotein. Indeed, NKp46–Ig binding was blocked when infected cells were first incubated with anti-HN mAbs TC-1D6 or TC-9C1 but not with the TC-9A1 mAb that is specific for the other principal SV glycoprotein—the fusion (F) protein (Table 1).
Despite the dependency on HN of NKp46–Ig binding to SV-infected 721.221 cells, we failed to detect an increased susceptibility of these cells to NK-mediated lysis associated with SV infection (data not shown), perhaps because non-infected 721.221 cells are already sensitive to NK-mediated lysis and/or receptors other than

Table 1 Anti-haemagglutinin antibodies inhibit NKp46–Ig binding to Sendai-virus-infected cells

mAb specificity	mAb binding (MFI)		NKp46–Ig binding (MFI)	
	721.221	721.221 SV	721.221	721.221 SV
No mAb	–	5	5	6
TC-9A1	Anti-Fusion	5	82	6
TC-1D6	Anti-HN	7	141	7
TC-9C1	Anti-HN	13	179	7
135.7	Anti-HN	5	95	6

721.221 cells (10⁶ per ml) were incubated overnight with 100 μl ml⁻¹ of SV-containing supernatant. Cells (infected or uninfected) were washed, incubated with the indicated mAbs and stained either with FITC-labelled goat anti-mouse immunoglobulin, or with NKp46–Ig fusion protein followed by PE-conjugated goat anti-human Fc. MFI, median fluorescence intensity; the MFI was rounded to the nearest whole number. Background staining of SV-infected 721.221 and 721.221 cells with the PE-conjugated anti-human Fc was 4 and 2, respectively. Results are representative of five independent experiments. Similar results were obtained with the B-cell line RPMI 8866.

NKp46 are dominant in triggering the lysis of 721.221. Thus, to test the role of NKp46 recognition of HN in NK lysis, we transiently transfected 293T cells with a plasmid encoding HN. Forty-eight hours after transfection, cell-surface HN expression was confirmed using mAbs TC-1D6 (Fig. 1a) or 135.7 (data not shown). Notably, transfection resulted in a twofold increase in NKp46–Ig staining, without enhancing the staining with other immunoglobulin fusion proteins, KIR-1–Ig, NKAT-8–Ig or CD16–Ig (Fig. 1a). Moreover, NK GAL, an NK line derived from healthy donor peripheral blood lymphocytes (PBLs), lysed HN-transfected 293T cells at least fourfold more efficiently than non-transfected or mock-transfected cells (Fig. 1b, c). Pre-incubation of NK GAL with a mouse antiserum raised against NKp46–Ig inhibited the increased killing of HN-transfected 293T cells, whereas incubation with a control serum had little effect (Fig. 1b). The same experiment revealed that each of the three HN-specific mAbs could block NK-mediated lysis, but a control mAb specific for CD99 had no significant effect (Fig. 1c).
One of the hallmarks of NK recognition is its lack of antigen specificity. Having shown that NKp46 interacts with the SV HN in NK-mediated lysis, we next tested whether it could interact with the influenza virus (IV) haemagglutinin (HA). IV infection of 1106mel cells (a class-I-deficient cell line) resulted in a fourfold increase in NKp46–Ig binding (Table 2). As above, the specific nature of the enhanced binding is shown by the constant binding of other immunoglobulin fusion proteins (data not shown). Notably, the increased NKp46–Ig binding was completely or partially blocked, respectively, by the HA-specific mAbs H28-E23 or H17-L2 (Table 2). In contrast, HN-specific mAbs TC-1D6, TC-9C1 or 135.7 had no effect on binding (data not shown).
We next showed that IV infection of 1106mel cells enhances NK-GAL-mediated lysis, and that the enhanced killing is blocked

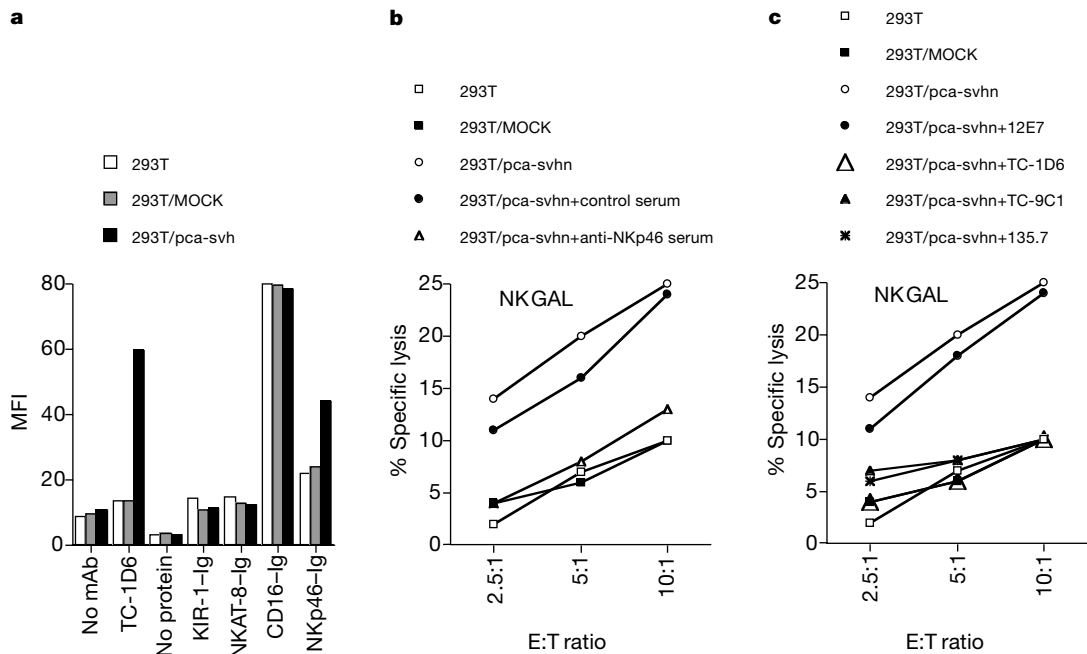


Figure 1 Lysis of HN-transfected 293T cells by NK GAL. **a**, NKp46–Ig binding to 293T cells transfected with Sendai virus haemagglutinin cDNA. 293T cells were transiently transfected with a control PCDNA3 plasmid (293T/MOCK) or with a cDNA encoding for SV HN (293T/pca-svhn) using Fugene reagent (Boehringer Mannheim). After 48 h, cells were stained either with TC-1D6 mAb or with KIR-1, NKAT-8, CD16 and NKp46 immunoglobulin fusion proteins. MFI, median fluorescence intensity. Controls were the same cells stained either with FITC-conjugated anti-mouse antibodies (no mAb), or with PE-conjugated anti-human Fc antibodies (no protein). Results represent one experiment of three performed. **b**, **c**, Enhanced lysis of 293T cells transfected with the Sendai virus

HN cDNA is blocked by anti-NKp46 and anti-HN mAb. Forty-eight hours after transfection, 293T, 293T/MOCK and 293T/pca-svhn cells were labelled with [³⁵S]methionine and washed. Cells were then incubated at the effector to target (E:T) ratios indicated; with NK GAL pre-incubated with either control serum or with anti-NKp46 serum (**b**). Alternatively, cells were incubated with the indicated mAb for 1 h on ice, washed, and incubated with NK GAL (**c**). In all experiments, NK cells were pre-incubated with 50% human serum for 1 h on ice and washed to block Fc receptors. Results represent one experiments of three performed in both **b** and **c**.

by pre-incubation of cells with anti-NKp46 serum but not with control serum or mAbs 12E7 and 3G8, which are specific for CD99 and CD16, respectively (Fig. 2a). Incubation of IV-infected 1106mel cells with mAb H28-E23 resulted in complete inhibition of the increased lysis, whereas mAb H17-L2 had a partial inhibitory effect, mirroring the results of the blocking experiment (Fig. 2b, Table 2).

We also examined the recognition of IV-infected 1106mel cells by clones prepared from NK GAL by limiting dilution. All 28 clones tested were positively stained with the anti-NKp46 serum. Twenty-one of the clones exhibited enhanced recognition of IV-infected cells relative to uninfected cells (Fig. 2c); pre-incubation of IV-infected 1106mel cells with the HA-specific mAb H28-E23 completely inhibited the IV-enhanced lysis by all 21 NK clones (data not shown). Pre-incubation of NK clones with anti-NKp46 serum largely inhibited the IV-enhanced lysis of 13 of these clones, (for example clone 6, Fig. 2c), whereas 6 of the clones were partially inhibited (for example clone 15, Fig. 2c). The average median fluorescence intensity (MFI) (median fluorescence intensity) after staining with anti-NKp46 serum for these two groups was 28.9 and 32.8, respectively.

Table 2 Anti-haemagglutinin antibodies inhibit NKp46-Ig binding to influenza-virus-infected cells

mAb specificity		mAb binding (MFI)		NKp46-Ig binding (MFI)	
		1106mel	1106mel IV	1106mel	1106mel IV
No mAb	—	18	22	112	441
H28-E23	Anti-HA	18	2,016	100	63
H17-L2	Anti-HA	11	2,100	110	145
NA2-1C1	Anti-NA	10	2,000	109	980

1106mel cells (10^6 per ml) were incubated overnight with $1,000 \text{ U ml}^{-1}$ of IV. Cells (infected or uninfected) were washed, incubated with the indicated mAbs, and stained with either FITC-conjugated goat anti-mouse immunoglobulin or the NKp46-Ig fusion protein followed by PE-conjugated goat anti-human Fc. MFI, median fluorescence intensity; the MFI was rounded to the nearest whole number. Background staining of IV-infected 1106mel and 1106mel cells with the PE-conjugated anti-human Fc was 7 and 6, respectively. Results are representative of eight independent experiments

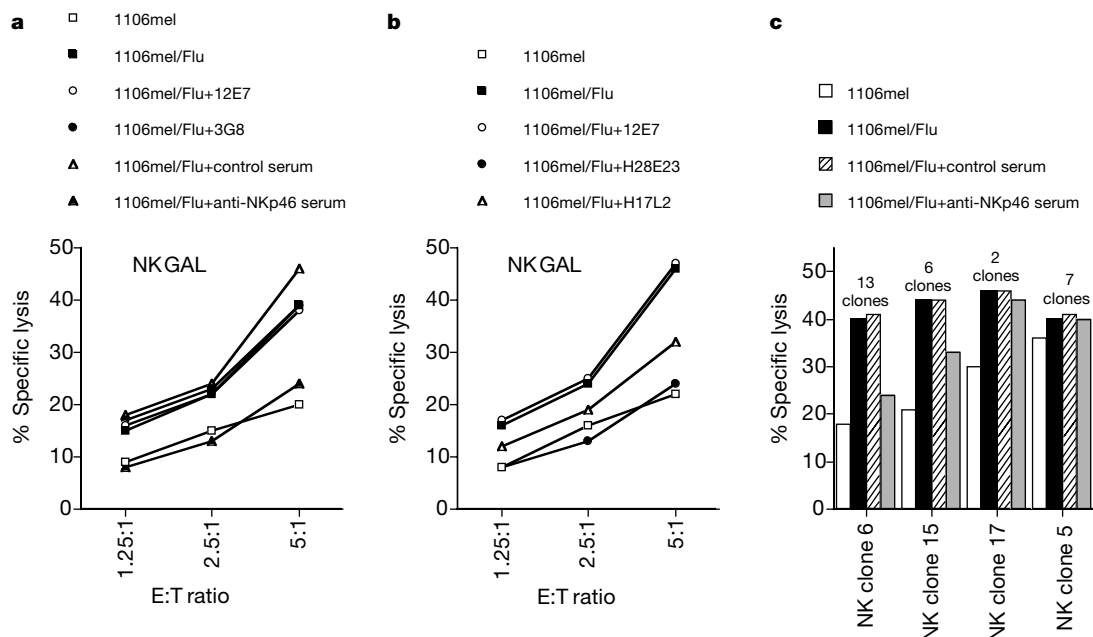


Figure 2 Lysis of IV-infected 1106mel cells by NK GAL and derived clones. **a**, NK cells were incubated with no antibody, control anti-CD99 mAb (12E7), anti-CD16 mAb (3G8), control serum, or anti-NKp46 serum for 1 h on ice. NK cells were then washed and incubated with either 1106mel or IV-infected 1106mel cells at the indicated E:T ratios. **b**, IV-infected 1106mel cells were incubated with various mAbs for 1 h on ice, and then with NK GAL at the indicated E:T ratios. Results represent one experiment of three

Two clones that showed enhanced lysis of IV-infected cells were not affected by the NKp46 antiserum (for example clone 17, Fig. 2c). The average MFI for anti-NKp46 staining for this group of two clones was 17.5. Finally, no IV-associated enhancement in lysis was observed in seven of the clones tested (for example clone 5, Fig. 2c), and the MFI for anti-NKp46 staining of this group was 31.3. Similar results were obtained when the same NK clones were tested with HN-transfected 293T cells (data not shown). These findings indicate, first, that NKp46 is required for the recognition of HA- and HN-expressing cells by a substantial subset of NK cells; and second, that other populations of NK cells can lyse these cells in an NKp46-independent manner.

Previous reports indicated that haemagglutinin can enhance lysis of virus-infected target cells by NK cells, as well as activate NK cells directly^{1,16,17}. Our observations argue that this is mediated by interaction with the triggering receptor NKp46, rather than by a nonspecifically increased conjugation effect. NKp46-Ig binds to 293T transfectants expressing HN (Fig. 1a), as well as to IV-infected 1106mel cells (Table 2). The enhancement in binding of NKp46-Ig to 1106mel after IV infection can be blocked with purified HA (Fig. 3a). In addition, purified HA binds to peripheral blood NK cells, and this binding can be blocked by NKp46-Ig but not by control CD99-Ig (data not shown). These data suggest that there is a direct interaction between NKp46 and haemagglutinin.

To test further this direct activation through the interaction of HA and cellular NKp46, we transfected BW cells with constructs encoding chimaeric proteins in which extracellular portion of the NKp46 (BW-NKp46) or CD16 (BW-CD16) are fused to the transmembrane and tail domains of the CD3 ζ chain. We incubated the transfectants or untransfected control cells with influenza virus for 20 h and tested supernatants for the presence of interleukin (IL)-2. This revealed that virus induced IL-2 secretion from BW-NKp46 but not from BW-CD16 or untransfected BW cells (Fig. 3b). This suggests that HA, presented in a multivalent state on IV, can ligate NKp46 in such a manner that may activate effector function.

performed in both **a** and **b**. **c**, Twenty-eight NK clones were derived from the NK GAL by limiting dilution. Blocking experiments with serum containing polyclonal antibodies were as in Fig. 1. The E:T ratio was 5:1. Numbers indicate the number of clones that behaved similarly to the one NK clone presented. Results represent one experiment of two performed. In all experiments, NK cells were pre-incubated with 50% human serum for 1 h on ice and washed to block Fc receptors.

Thus, although on a molar basis NKp46 may not be the chief haemagglutinin receptor on NK cells, it can be a functional haemagglutinin receptor.

Sendai virus HN and influenza virus HA both recognize terminal N-acetylneuraminic acid residues (sialic acids) attached to galactose, suggesting that they use a common mechanism for binding to NKp46. The involvement of sialic acid in the interaction of NKp46 with HA is indicated by a number of findings.

First, NKp46-Ig can completely block IV-mediated agglutination of sheep erythrocytes at a protein concentration as low as 2 μ M (data not shown). Second, pre-incubation of IV-infected cells with a mAb (NA2-1C21) that blocks the enzymatic activity of IV neuraminidase (NA, the other principal IV glycoprotein expressed on the surface of infected cells and virions) significantly enhances NKp46-Ig binding to IV-infected cells (Table 2). Third, NKp46-Ig does not stain target cells infected with measles (data not shown) whose HA does not bind sialic acid¹⁸. Last, and most directly, treatment of NKp46-Ig with bacterial neuraminidase reduced its binding to IV-infected cells without reducing its binding to uninfected cells (Fig. 3c).

Treatment of NKp46-Ig did not effect staining of non-infected 1106mel cells, measured by flow cytometry (Fig. 3c), nor alter the integrity of the protein tested by SDS-PAGE analysis (data not

shown). As de-sialylation often increases interactions by reducing negative charge repulsion of receptor–ligand pairs¹⁹, this strongly supports the direct interaction of NKp46 with the sialic-binding site of HA. Haemagglutinins of different influenza viruses have different specificities for sialic acid linked to galactose by α 2,3 or α 2,6 linkages²⁰, both expressed by NK cells^{21,22}.

We interpret these findings to indicate that NKp46 binds to target cells in two ways: first, through the interaction of NKp46-associated sialic acid with viral sialic acid receptors; second, in a sialic-acid-independent interaction with undefined cellular ligands. The former is clearly responsible for the enhanced killing of IV-infected cells by the NK cells that we have studied. It cannot be absolutely ruled out that the interaction of HA with NKp46 could principally augment NK cytotoxicity by facilitating binding of NKp46 to a second cellular receptor. It is also unclear whether the interaction of NKp46 with viral HA is sufficient for triggering lysis. The role of non-viral ligands to NKp46 and the role of other NK-triggering receptors, for recognition of viral infected cells, is clearly an important area of further research.

The dissociation constant of HA for sialic acids is in the millimolar range—too low for a monomeric interaction to account for either the stable binding of NKp46-Ig detected by flow cytometry (which requires dissociation constants less than 0.1 μ M), or the

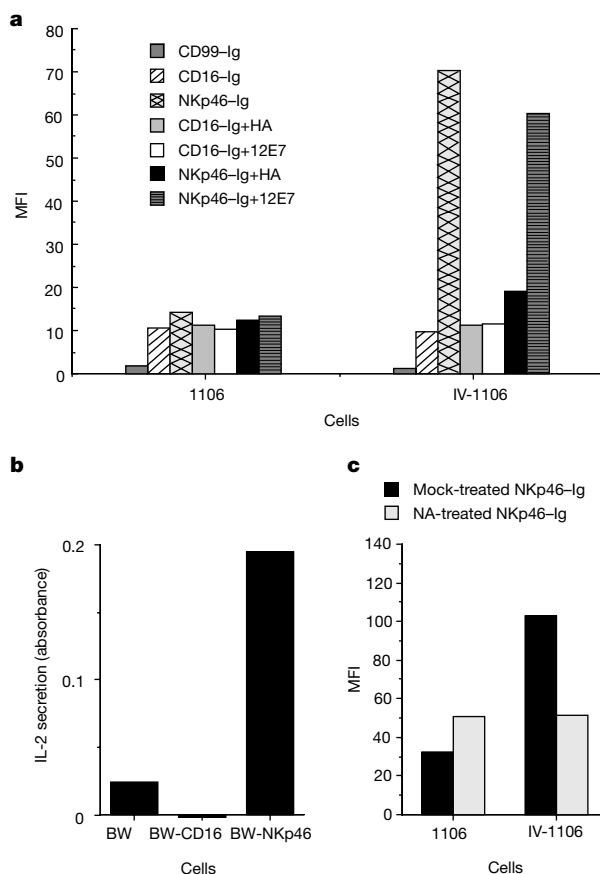


Figure 3 NKp46 binding to and activation by haemagglutinin. **a**, NKp46-Ig was incubated with or without 40 μ g purified HA protein (no blocking of NKp46-Ig binding was evident when less than 40 μ g of purified HA protein was used). Mixtures were next incubated with IV-infected or non-infected 1106mel cells and stained with PE-conjugated goat anti-human Fc. Results represent one experiment of two performed. MFI, median fluorescence intensity. Similar results were obtained when SV-infected 721.221 cells were used. **b**, Cells (10^6 in 1 ml) were incubated with 1,000 U IV for 20 h in 24-well plates. IL-2 levels in the supernatants were measured by ELISA, and results are given as absorbance (optical density 650). Absorbances from ELISAs of supernatants of untreated cells (0.144, 0.127 and 0.238 for BW, BW-CD16 and BW-NKp46, respectively) were

subtracted from those of treated cells. BW cells and transfectants were not susceptible to IV infection in these conditions, as shown by the negative staining of cells with anti-NA mAb. Results represent one experiment of three performed. **c**, NKp46-Ig was incubated with 0.01 U of insoluble neuraminidase attached to beaded agarose (N-5254; Sigma) or with PBS (control) for 1 h at 17 °C on a roller. IV-infected, or non-infected 1106mel cells were washed, and stained either with NA-treated or mock-treated NKp46-Ig, followed by PE-conjugated anti-human Fc. Results represent one experiment of eight performed. The activity of NA was confirmed by SDS-PAGE analysis of NA-treated fetuin, a highly sialylated protein.

potency of NKp46-Ig in blocking viral haemagglutination. This implies either a multimeric interaction of HA with NKp46 or that NKp46 interacts more intimately with HA after contact is initiated by the sialic-acid binding. Even in the former case, NKp46 would need special properties to distinguish it from other cellular glycoproteins, as terminal sialic residues are ubiquitous on N- and O-linked oligosaccharides found on glycoproteins. One possibility is that NKp46 (which is thought to have a single N-linked and two O-linked oligosaccharides¹⁴) oligomerizes at the cell surface to enable multivalent interaction of sialic acid with a single HA complex, which as a trimer possesses three sialic-acid-binding sites.

The existence of NK clones that recognize IV-infected cells in an NKp46-independent manner (for example NK GAL clone 17, Fig. 2c) infers that other lysis receptors are involved in the recognition of virus-infected cells (for example NKp44; our unpublished data). Other receptors may also recognize HA, as the enhanced lysis associated with virus infection can be completely blocked by anti-HA mAb. Given that members of at least seven virus families use sialic acid as a receptor for virus entry into host cells, this suggests a general strategy for NK cell recognition of a substantial subset of viruses. □

Methods

Cells and viruses

We used the following cell lines: the MHC class-I-negative human EBV-transformed B-cell line 721.221; the MHC class-I-negative human melanoma line 1106mel; and the adenovirus-transformed, SV40-large T-antigen-transfected, human fibroblast kidney cell line 293T. NK cells (lines and clones) were isolated from PBLs using the human NK cell isolation kit and the autoMACS instrument (Miltenyi Biotec Inc); NK cells were kept in culture as described²³. SV and the IV A/PR/8/34 (H1N1) were purchased from Spafas (Preston City, CT, USA).

Monoclonal antibodies

SV-specific mAb^{24,25} and IV-specific mAb²⁶ have been described. Anti-CD99 mAb 12E7 was a gift from A. Bernard (Hopital de L'Archet, Nice, France). The hybridoma-producing mAb 3G8 was given by J. Unkeless (Mt Sinai School of Medicine, New York, USA).

Cytotoxicity assays

The cytotoxic activity of NK lines and clones against the various targets was assessed in 5-h ³⁵S-release assays²⁷. In experiments where mAbs were included, NK cells were first incubated with 50% human serum (to prevent binding of the mAb to the various Fc receptors expressed on the surface of human NK cells) and washed. The final mAb concentration was 20 µg ml⁻¹, or a 1:100 dilution in cases where the mAbs were present in ascites from hybridoma-bearing mice. In all experiments, the spontaneous release was less than 25% of maximal release. Each point represents the average of duplicate values. The range of the duplicates was within 5% of their mean.

Immunoglobulin fusion proteins

The generation of CD16-Ig fusion protein has been described¹⁰. DNA fragments encoding the extracellular portions of NKp46 (isoform accession number AJ006121), KIR-1 (accession number L41267) and NKAT-8 (NM_012314) were amplified by polymerase chain reaction (PCR) from complementary DNA isolated from human NK clones. These PCR-generated fragments were cloned into the mammalian expression vector containing the Fc portion of human IgG1, and immunoglobulin fusion proteins were produced as described¹⁰. Sequencing of the constructs revealed that the cDNAs of all immunoglobulin fusion proteins were in frame with the human Fc genomic DNA and were identical to the reported sequences. All immunoglobulin fusion proteins used in this work migrate as a single band on standard non-reduced SDS-PAGE gels, and each was regularly assayed by SDS-PAGE to ensure that the proteins had not degraded. The procedure for staining cells with immunoglobulin fusion proteins has been described¹⁰.

Production of anti-NKp46 serum

BALB/c mice were injected in their foot pads with 40 µg of NKp46-Ig or KIR-1-Ig fusion proteins emulsified in complete Freund's adjuvant (CFA). Mice were boosted after 6 weeks, and sera were collected 12 d later. For control sera, mice were immunized as above with PBS emulsified in CFA. Sera were tested for NKp46 antigen specificity on YTS, KIR-1-transfected YTS cells, various NK lines and clones, and non-NK lines. The anti-NKp46 serum was used at a 1:100 dilution, a concentration at which binding was saturated, as measured by flow cytometry. In agreement with the pattern of expression of NKp46 reported¹⁴, only NK cells were positively stained with the anti-NKp46 serum, and control cells 721.221, RPMI 8866, Jurkat and others remained unstained. All NK cells (4 lines and more than 50 clones) were stained to various degrees with this serum. Most directly, sera

were tested on NKp46-, CD16- and CD99-stable transfectants of BW cells, and data clearly show the specificity of the anti-NKp46 serum. In addition, re-directed lysis of P815 cells could be induced when the cells were coated with the anti-NKp46 serum and incubated with various NK lines and clones.

Blocking of NKp46-Ig binding to virus-infected cells

Cells (721.221 and 1106mel) were infected either with 100 µl of SV supernatant (for 721.221), or with 1,000 units ml⁻¹ of IV (for 1106mel). After overnight incubation, infected cells were washed and incubated with the various mAbs for 1 h on ice. Next, cells were washed and assayed for staining with 10 µg of the appropriate immunoglobulin fusion protein as described¹⁰.

Blocking of NKp46-Ig binding using purified HA

Ten micrograms of various immunoglobulin fusion proteins were incubated with 40 µg of purified HA protein²⁸ at final volume of 100 µl PBS/BSA/azide, for 2 h on ice. These mixtures were then incubated with cells for 2 h on ice and stained for the presence of immunoglobulin fusion proteins using described staining procedures¹⁰.

Activation of NKp46-CD3ζ chimaera

Murine BW thymoma cells (TCRαβ negative) were transfected with constructs encoding chimaeric proteins in which the extracellular portion of NKp46 or CD16 is fused to the transmembrane and tail domains of the CD3ζ chain. The cytoplasmic domain of the T-cell receptor ζ-chain is sufficient to trigger IL-2 secretion on ligation of extracellular portion²⁹. BW and stable BW transfectants were incubated with influenza virus for 20 h, and cell supernatants were assayed for the presence of IL-2 by ELISA kit (PharMingen, San Diego, CA).

Received 20 October; accepted 5 December 2000.

- Trinchieri, G. Biology of natural killer cells. *Adv. Immunol.* **47**, 187–376 (1989).
- Lanier, L. L. NK cell receptors. *Annu. Rev. Immunol.* **16**, 359–393 (1998).
- Ljunggren, H. G. & Karre, K. In search of the 'missing self': MHC molecules and NK recognition. *Immunol. Today* **11**, 7–10 (1990).
- Davis, D. M. *et al.* The human natural killer cell immune synapse. *Proc. Natl Acad. Sci. USA* **96**, 15062–15067 (1999).
- Braud, V. M. *et al.* HLA-E binds to natural killer cell receptors CD94/NKG2A, B and C. *Nature* **391**, 795–799 (1998).
- Colonna, M. & Samaridis, J. Cloning of immunoglobulin-superfamily members associated with HLA-C and HLA-B recognition by human natural killer cells. *Science* **268**, 405–408 (1995).
- Wagtmann, N. *et al.* Molecular clones of the p58 NK cell receptor reveal immunoglobulin-related molecules with diversity in both the extra- and intracellular domains. *Immunity* **2**, 439–449 (1995).
- D'Andrea, A. *et al.* Molecular cloning of NKBI. A natural killer cell receptor for HLA-B allotypes. *J. Immunol.* **155**, 2306–2310 (1995).
- Bottino, C., Biondi, R., Millo, R., Moretta, L. & Moretta, A. The human natural cytotoxicity receptors (NCR) that induce HLA class I independent NK cell triggering. *Hum. Immunol.* **61**, 1–6 (2000).
- Mandelboim, O. *et al.* Human CD16 as a lysis receptor mediating direct natural killer (NK) cell cytotoxicity. *Proc. Natl Acad. Sci. USA* **96**, 5640–5644 (1999).
- Pende, D. *et al.* Identification and molecular characterization of NKp30, a novel triggering receptor involved in natural cytotoxicity mediated by human natural killer cells. *J. Exp. Med.* **190**, 1505–1516 (1999).
- Cantoni, C. *et al.* NKp44, A triggering receptor involved in tumor cell lysis by activated human natural killer cells, is a novel member of the immunoglobulin superfamily. *J. Exp. Med.* **189**, 787–796 (1999).
- Biondi, R. *et al.* The murine homologue of the human NKp46, a triggering receptor involved in the induction of natural cytotoxicity. *Eur. J. Immunol.* **29**, 1014–1020 (1999).
- Pessino, A. *et al.* Molecular cloning of NKp46: novel member of the immunoglobulin superfamily involved in triggering of natural cytotoxicity. *J. Exp. Med.* **188**, 953–960 (1998).
- Sivori, S. *et al.* NKp46 is the major triggering receptor involved in the natural cytotoxicity of fresh or cultured human NK cells. Correlation between surface density of NKp46 and natural cytotoxicity against autologous, allogeneic or xenogeneic target cells. *Eur. J. Immunol.* **29**, 1656–1666 (1999).
- Alsheikhly, A. *et al.* Sendai-virus-induced cell-mediated cytotoxicity *in vitro*. The role of viral glycoproteins in cell-mediated cytotoxicity. *Scand. J. Immunol.* **17**, 129–138 (1983).
- Alsheikhly, A. R., Orvell, C., Andersson, T. & Perlmann, P. The role of serologically defined epitopes on mumps virus HN-glycoprotein in the induction of virus-dependent cell-mediated cytotoxicity. Analysis with monoclonal antibodies. *Scand. J. Immunol.* **22**, 529–538 (1985).
- Maisner, A. & Herrler, G. Membrane cofactor protein with different types of N-glycans can serve as measles virus receptor. *Virology* **210**, 479–481 (1995).
- Varki, A. Sialic acids as ligands in recognition phenomena. *FASEB J.* **11**, 248–255 (1997).
- Weis, W. *et al.* Structure of the influenza virus hemagglutinin complexed with its receptor, sialic acid. *Nature* **333**, 426–431 (1988).
- De Lau, W. B., Kuipers, J., Voshol, H., Clevers, H. & Bast, B. J. HB4 antibody recognizes a carbohydrate structure on lymphocyte surface proteins related to HB6, CDw75, and CD76 antigens. *J. Immunol.* **150**, 4911–4919 (1993).
- Pinola, M., Renkonen, R., Majuri, M. L., Tiisala, S. & Saksela, E. Characterization of the E-selectin ligand on NK cells. *J. Immunol.* **152**, 3586–3594 (1994).
- Mandelboim, O. *et al.* Protection from lysis by natural killer cells of group 1 and 2 specificity is mediated by residue 80 in human histocompatibility leukocyte antigen C alleles and also occurs with empty major histocompatibility complex molecules. *J. Exp. Med.* **184**, 913–922 (1996).
- Peterhans, E., Bachi, T. & Yewdell, J. W. Evidence for different receptor sites in mouse spleen cells for the SV hemagglutinin-neuraminidase (HN) and fusion (F) glycoproteins. *Virology* **128**, 366–376 (1983).
- Yewdell, J. W. & Gerhard, W. Delineation of four antigenic sites on a paramyxovirus glycoprotein via which antibodies mediate distinct antiviral activities. *J. Immunol.* **128**, 2670–2675 (1982).

26. Yewdell, J. W., Gerhard, W. & Bachi, T. Monoclonal anti-hemagglutinin antibodies detect irreversible antigenic alterations that coincide with the acid activation of influenza virus A/PR/8/34-mediated hemolysis. *J. Virol.* **48**, 239–248 (1983).
27. Porgador, A., Mandelboim, O., Restifo, N. P., Strominger, J. L. Natural killer cell lines kill autologous beta2-microglobulin-deficient melanoma cells: implications for cancer immunotherapy. *Proc. Natl Acad. Sci. USA* **94**, 13140–13145 (1997).
28. Brand, C. M. & Skehel, J. J. Crystalline antigen from the influenza virus envelope. *Nature* **238**, 145–147 (1972).
29. Irving, B. A. & Weiss, A. The cytoplasmic domain of the T cell receptor zeta chain is sufficient to couple to receptor-associated signal transduction pathways. *Cell* **64**, 891–901 (1991).

Acknowledgements

We thank N. P. Restifo for the 1106mel cells; A. Portner from St. Jude Children's Research Hospital for providing the pca-svhn plasmid; and J. Skehel and R. Gonsalves from National Institute of Medical Research for the purified HA. O.M. is supported by the Israel Science Foundation and The Charles H. Revson Foundation. A.P. is supported by the Israel Science Foundation and by the USA–Israel Binational Science Foundation. O.M. and A.P. are supported by the Israel Cancer Research Foundation.

Correspondence and requests for materials should be addressed to A.P.
(e-mail: angel@bgumail.bgu.ac.il).

Identification of CRE1 as a cytokinin receptor from *Arabidopsis*

Tsutomu Inoue*, Masayuki Higuchi*, Yukari Hashimoto*, Motoaki Seki†, Masatomo Kobayashi†, Tomohiko Kato‡, Satoshi Tabata‡, Kazuo Shinozaki† & Tatsuo Kakimoto*

* Department of Biology, Graduate School of Science, Osaka University, Toyonaka, Osaka 560-0043, Japan

† Laboratory of Plant Molecular Biology, RIKEN Tsukuba Institute, 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan

‡ Kazusa DNA Research Institute, Kisarazu, Chiba 292-0812, Japan

Cytokinins are a class of plant hormones that are central to the regulation of cell division and differentiation in plants^{1,2}. It has been proposed that they are detected by a two-component system, because overexpression of the histidine kinase gene *CKI1* induces typical cytokinin responses³ and genes for a set of response regulators of two-component systems can be induced by cytokinins^{4,5}. Two-component systems use a histidine kinase as an environmental sensor and rely on a phosphorelay for signal transduction. They are common in microorganisms, and are also emerging as important signal detection routes in plants^{6–9}. Here we report the identification of a cytokinin receptor. We identified *Arabidopsis cre1* (cytokinin response 1) mutants, which exhibited reduced responses to cytokinins. The mutated gene *CRE1* encodes a histidine kinase. *CRE1* expression conferred a cytokinin-dependent growth phenotype on a yeast mutant that lacked the endogenous histidine kinase *SLN1* (ref. 10), providing direct evidence that *CRE1* is a cytokinin receptor. We also provide evidence that cytokinins can activate *CRE1* to initiate phosphorelay signalling.

Generally, cytokinins induce cell division, chloroplast development and formation of shoots (buds)¹. We screened mutagenized *Arabidopsis* for mutants that were impaired in cytokinin responses, including rapid cell proliferation and shoot formation in tissue culture. We isolated a mutant designated *cytokinin response 1-1* (*cre1-1*). We tested the responses of *cre1-1* to auxin and cytokinin in tissue culture, using naphthalene acetic acid (NAA) as an auxin and kinetin as a cytokinin (Fig. 1). Wild-type explants responded to increasing levels of kinetin with rapid proliferation, greening and formation of shoots (Fig. 1a). By contrast, such cytokinin responses were not evident in *cre1-1* (Fig. 1b). The mutant was also less

responsive to other cytokinins, including *trans*-zeatin, isopentenyladenine, benzyl adenine and the phenylurea-type synthetic cytokinin thidiazuron (see Supplementary Information).

Next we tested the responses of *cre1-1* to various plant hormones in a root elongation assay. External application of cytokinins¹¹, ethylene¹², auxins¹³ or abscisic acid¹⁴ inhibits root elongation. The root of the *cre1-1* mutant was less sensitive to benzyl adenine than that of wild-type plants, but it responded normally to the ethylene precursor 1-aminocyclopropane-1-carboxylic acid (ACC) and the auxin indole-3-acetic acid (IAA) (Fig. 2a–c). The responses of *cre1-1* to low levels of abscisic acid (ABA) were slightly higher than normal (Fig. 2d). The cytokinin responses of *cre1-1* heterozygotes were intermediate between those of *cre1-1* homozygotes and the wild type (see Supplementary Information).

We mapped the *CRE1* locus to the top of chromosome 2 between the *rga* and *ngal145* markers (see Supplementary Information). We searched the genome sequence of *Arabidopsis* between these markers for genes that could code for proteins involved in signal transduction. Among them was the hypothetical gene *At2g01830*, possibly coding for a histidine kinase. The nucleotide sequence of *At2g01830* revealed that this gene was mutated in the *cre1-1* mutant. Hereafter we refer to this gene as *CRE1*. *CRE1* is identical to *WOL*¹⁵ (see below) and *AHK4* (ref. 16). A genomic fragment containing *CRE1* was introduced into *cre1-1* mutant calli. Wild-type calli that had been transformed with the control vector regenerated shoots when cultured in the presence of the

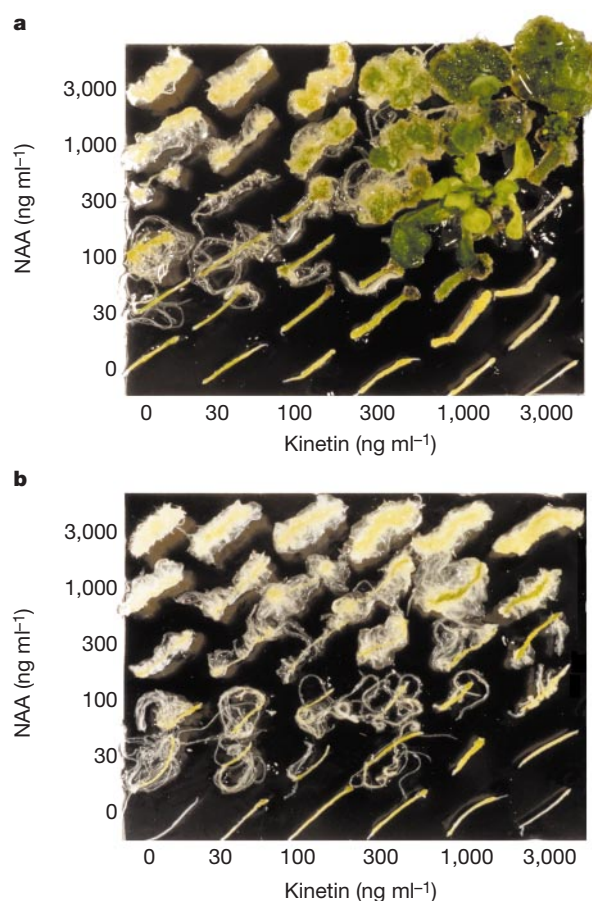


Figure 1 Callus growth of the cytokinin-resistant mutant *cre1-1* in different auxin and cytokinin concentrations. Hypocotyl segments were excised and cultured on media containing different levels of kinetin and NAA. After 21 days in culture, the induced calli were arranged and photographed. Wild-type explants (**a**) proliferated rapidly, turned green, and produced shoots in the presence of high concentrations of cytokinins. The *cre1-1* explants (**b**) did not.

cytokinin *trans*-zeatin, but *cre1-1* mutant calli transformed with the control vector did not (Fig. 3). However, mutant calli regenerated shoots in the presence of *trans*-zeatin if they had been transformed with pGPTV-KAN-CRE1 (Fig. 3d), indicating that CRE1 complemented the *cre1-1* mutant.

We screened a library of *Arabidopsis* complementary DNA and isolated the corresponding cDNA clones, which were derived from two types of alternately spliced message named CRE1a and CRE1b. The predicted protein for CRE1a consists of 1,057 amino acids and that for CRE1b has 23 extra amino acids at its amino terminus. The CRE1a product was used for further study and for numbering of the amino-acid residues. The carboxy-terminal region of CRE1 carries a histidine kinase domain and a receiver domain. Between these domains, there is another region with weak similarity to receiver domains. According to the PSORT prediction (<http://psort.nibb.ac.jp/form.html>), CRE1 probably localizes to the plasma membrane. The N-terminal region probably consists of an extracellular domain flanked by two transmembrane segments, and the C-terminal region is intracellular. We detected CRE1 message in various tissues (data not shown), but the highest expression was in the root¹⁵. The *cre1-1* mutation converted Gly 467 in the histidine kinase domain to Asp 467 (see Supplementary Information). *Arabidopsis* has genes for two products, AAF99730 (AHK3 (ref. 16) and BAB09274 (AHK2 (ref. 16)), that share high sequence similarity to CRE1, being 52% and 54% identical, respectively, over their entire proteins, and 61% and 60% identical, respectively, over their extracellular domains. CKI1 was less similar to CRE1 than these proteins (see Supplementary Information).

We also isolated an *Arabidopsis* line, *cre1-2*, with a T-DNA (see Methods) insertion in the CRE1 gene. The integration occurred in the place of nine base pairs of CRE1 between nucleotide positions +75 and +84 relative to the inferred translation initiation site (see

Supplementary Information). The *cre1-2* line, which was homozygous for the T-DNA insertion, was also resistant specifically to cytokinins in the root elongation assay (Fig. 2) and in the callus growth and shoot formation assays (data not shown). The *cre1-2* mutant was complemented by introduction of the CRE1 gene (see Supplementary Information). The presence of the mutation in the CRE1 gene in either of the *cre1-1* or *cre1-2* mutants, and complementation of *cre1-1* and *cre1-2* by CRE1, are definitive evidence that mutations in CRE1 cause the cytokinin-insensitive phenotype of the *cre1* mutants. The *cre1* mutants were allelic to the *wol* mutant (that is, the *cre1* and *wol* mutants bore mutations in the same gene), which is impaired in cell division and proper formation of vascular tissue of the root¹⁵. The xylem organization of *cre1-1* was also altered (data not shown).

To explore the function of the CRE1 gene, we expressed CRE1 (Fig. 4) in a yeast strain deficient in the SLN1 gene, which encodes an osmosensing histidine kinase¹⁰. At normal osmolarity, SLN1 autophosphorylates the conserved histidine residue. The phosphoryl group is then transferred to the conserved aspartate residue in the receiver domain of the same protein, then to the phosphotransfer mediator YPD1, and finally to the SSK1 response regulator. This in turn inhibits the ability of SSK1 to activate the downstream mitogen-activated protein (MAP) kinase pathway¹⁷. The *sln1Δ* mutant is lethal because the downstream SSK1 is always dephosphorylated, which overactivates the downstream MAPK pathway^{10,17}. The *sln1Δ* mutant carrying p415CYC-CRE1, which would express the CRE1 gene, was still lethal without cytokinins. However, surprisingly, it grew at a normal rate if *trans*-zeatin, a native cytokinin, was included in the medium. It is noteworthy that the active cytokinin *trans*-zeatin¹⁸ was effective in this yeast system, but the much less active cytokinin *cis*-zeatin¹⁸ was ineffective (Fig. 4a). Other active cytokinins—*isopentenyladenine*, benzyl adenine and thidiazuron—were also effective. The plant hormones IAA, gibberellin A₃ and abscisic acid had no effect. Expression of CRE1b, another form of alternatively spliced product, in the *sln1Δ* mutant gave the same results (data not shown).

p415CYC-CRE1 did not suppress the lethality of the *ypd1Δ* mutant on plates either with or without cytokinins, indicating that

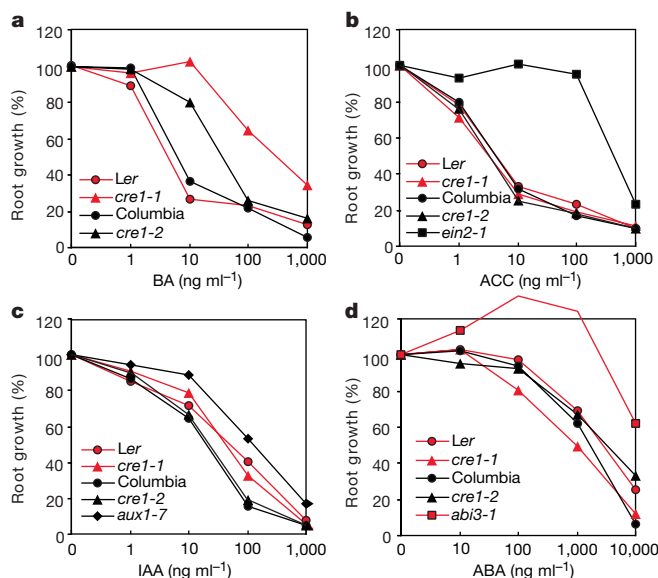


Figure 2 Root growth of the *cre1-1*, *cre1-2* and hormone-related mutants in the presence of various plant hormones. Seeds were sown on GM plates²² containing BA (a), ACC (b) or IAA (c). After chilling for three days, plates were incubated at 23 °C for eight days and the root lengths were measured. To determine ABA response, two-day-old seedlings germinated on GM were moved onto plates containing different concentrations of ABA and cultured for six days (d). Root growth was expressed relative to the mean root elongation of the same genotype on the medium without plant hormones. Each value represents the mean of at least 11 plants. *cre1-1* and *abi3-1* have Ler genetic background (red symbols); *cre1-2*, *ein2-1* and *aux1-7* have Columbia genetic background (black symbols). The root length of each genotype in the absence of plant hormones shown in a–d is given in the Supplementary Information.

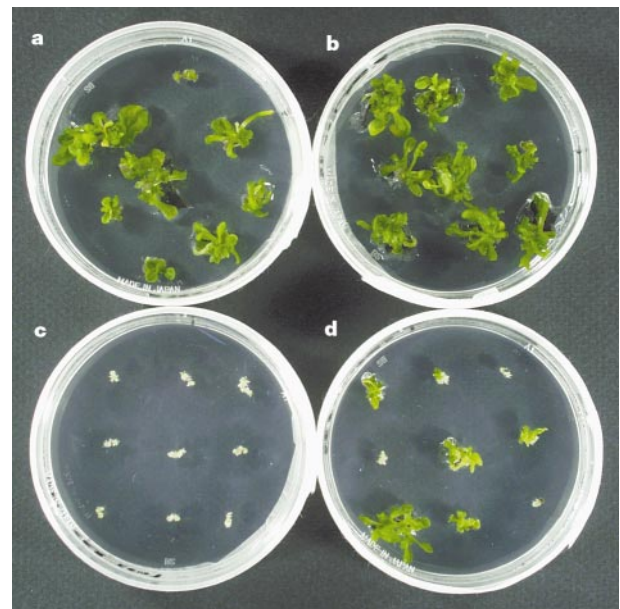


Figure 3 Complementation of *cre1-1* by CRE1. Wild-type calli (a, b) or *cre1-1* calli (c, d) were transformed with pGPTV-KAN (a, c) or pGPTV-KAN-CRE1 (b, d) and cultured for 19 days with 0.5 μg ml⁻¹ *trans*-zeatin and 0.3 μg ml⁻¹ indole butyric acid (an auxin). Shoots regenerated from different calli are independent transformants, and those on the same callus may or may not be independent.

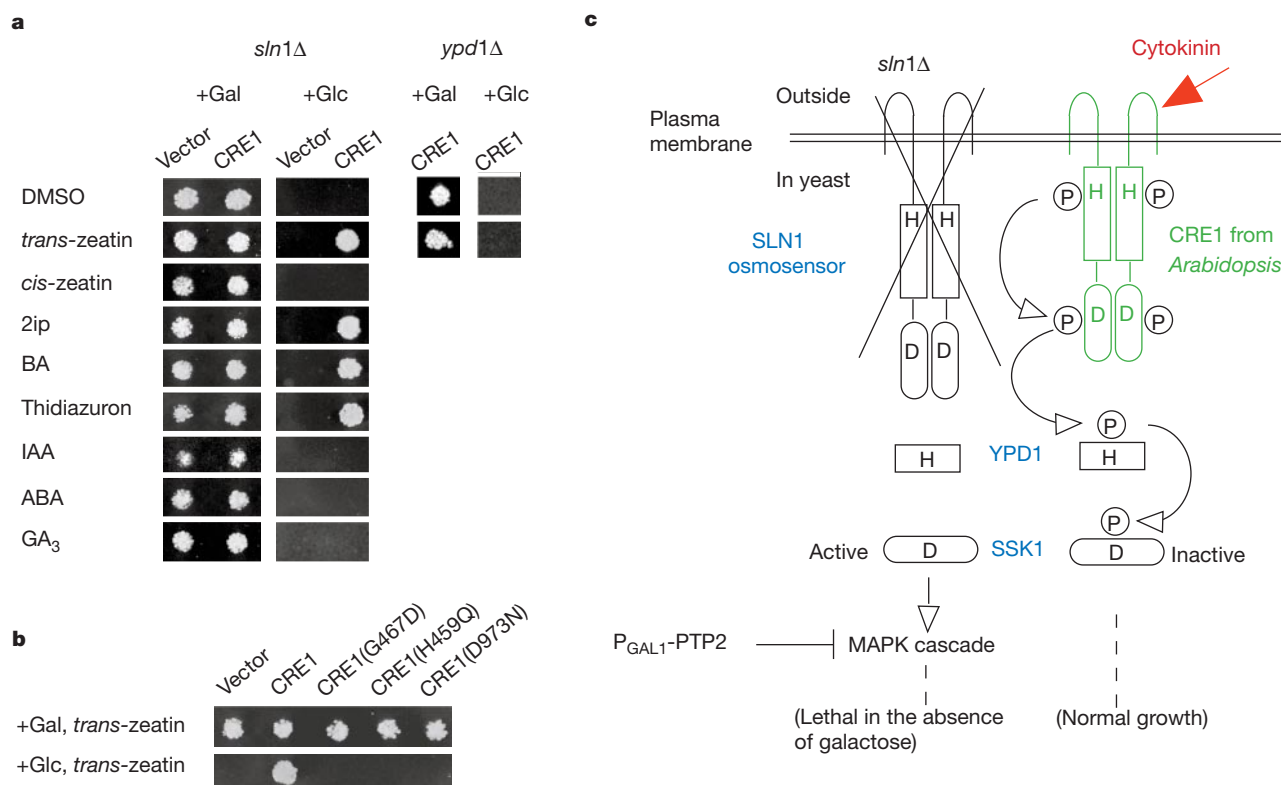


Figure 4 Cytokinin-dependent growth phenotype of yeast in which *SLN1* had been replaced with *CRE1*. **a**, *sln1Δ* and *ypd1Δ* strains were transformed with the vector p415CYC (vector) or p415CYC–*CRE1* (*CRE1*). Suspensions of transformants were spotted onto a plate containing a plant hormone as indicated, and galactose (+Gal) or glucose (+Glc). **b**, Effect of the mutation that was present in the *Arabidopsis cre1-1* mutant (G467D), or of the mutation that changed the conserved His 459 or Asp 973

phosphorylation sites (H459Q or D973N respectively). **c**, The presumed events in yeast. *CRE1* suppresses the lethality of *sln1Δ* but not *ypd1Δ* in the presence of cytokinins. The *sln1Δ* mutant is lethal because the dephosphorylated SSK1 constitutively activates the MAPK pathway. Cytokinins probably activate the histidine kinase activity of the *CRE1* protein to initiate the phosphorelay, whereby the phosphoryl group is transferred from the activated *CRE1* to *YPD1*, then to SSK1, suppressing the lethality of *sln1Δ*.

signal transduction from *CRE1* is mediated by *YPD1* in yeast (Fig. 4c). We next introduced the mutation that was present in the *Arabidopsis cre1-1* mutant into p415CYC–*CRE1*. The resulting plasmid, p415CYC–*CRE1*(G467D), could not suppress the lethality of *sln1Δ* either with or without *trans*-zeatin, indicating that the *CRE1* gene of the *Arabidopsis cre1-1* mutant was nonfunctional (Fig. 4b). Mutations at either the conserved His 459 or Asp 973 of the phosphorylation site in the histidine kinase or the receiver domains, respectively, also destroyed the ability of *CRE1* to suppress the lethality of the *sln1Δ* mutant (Fig. 4b). Therefore, cytokinins probably activate the histidine kinase activity of *CRE1*, and the signal is probably transmitted through *YPD1* to SSK1. *Arabidopsis* also has phosphotransfer mediators^{20,21}, which resemble *YPD1*, and response regulators^{4,5,7}. Therefore, in plants, cytokinins probably activate *CRE1* and possibly its homologues, which in turn initiate the phosphorelay signalling that governs cytokinin responses.

In *Arabidopsis*, ethylene receptors^{8,9} and possibly osmosensors¹⁹ are histidine kinases. CKI1 histidine kinase has been implicated in the detection or signal transduction (or both) of cytokinins³, but its function has yet to be clarified. We have provided evidence that the *CRE1* histidine kinase is a cytokinin receptor: mutations in the *CRE1* gene caused a cytokinin-insensitive phenotype in *Arabidopsis*, and expression of *CRE1* conferred a cytokinin-responsive phenotype on yeast. The *cre1* and *wol* mutants were impaired in the cell division and differentiation that is essential for proper formation of the root vascular tissue¹⁵. This observation, coupled with our data, underlines the importance of cytokinin signalling in this process. The *CRE1* homologues AAF99730 and BAB09274 may

also function as cytokinin receptors, which may explain why defects in *CRE1* did not cause more diverse phenotypes related to cytokinin functions. □

Methods

Screening for mutants impaired in cytokinin responses

Seeds of *A. thaliana* var. *Ler* were mutagenized with ethyl methanesulphonate, and seeds obtained after self-pollination (M2 seeds) were used²². Hypocotyl segments of M2 seedlings were aseptically excised and cultured on GM medium²² supplemented with 100 ng ml⁻¹ of 2,4-dichlorophenoxyacetic acid (2,4-D), 100 ng ml⁻¹ of kinetin, and vitamins (100, 10, 1, 1 and 1 μg ml⁻¹, respectively, of inositol, thiamine, nicotinic acid, pyridoxine HCl and biotin). Kinetin of this concentration is sufficient to induce cytokinin responses in wild-type explants, including induction of calli with rapid growth and greening without forming root primordia. The top portion of the plant, corresponding to each hypocotyl segment, was grown on GM medium. Calli with a reduced green colour with many root primordia, which usually occur under low levels of kinetin, were chosen as mutant candidates, and their seeds were obtained by growing the corresponding top portions. From about 19,000 M2 seedlings, one line was confirmed for the heritability of the callus phenotypes. Root growth was measured on GM medium supplemented with a plant hormone as described in Fig. 2.

Transformation of calli

A genomic region encompassing the *CRE1* gene was amplified by polymerase chain reaction (PCR) from genomic DNA of *Ler* using primers 5'-AGCACATGTGAGTTT-CACTGGCCTC-3' and 5'-AGCTCAAGTCGTCGACTGAGCTATAG-3'. The amplified fragment was digested with *Sall* and cloned into the pGPTV–KAN²³ vector between the *SmaI* and *Sall* sites. The sequence of the resultant construct, pGPTV–KAN–*CRE1*, was confirmed and was transformed into *Arabidopsis* calli by the *Agrobacterium*-mediated method, as described^{23,24}, except that hormone concentrations of CIM medium²⁴ were changed to 0.5 μg ml⁻¹ 2,4-D and 0.5 μg ml⁻¹ kinetin. The transformed calli were cultured on GM medium supplemented with 50 μg ml⁻¹ kanamycin sulphate, 100 μg ml⁻¹ cefotaxime, 100 μg ml⁻¹ vancomycin, 0.3 μg ml⁻¹ indole butyric acid and 0.5 μg ml⁻¹ *trans*-zeatin, and the same vitamins as were used in the medium for mutant screening. Other culture conditions were as described²¹.

Screening for T-DNA insertion lines

We used the T-DNA insertion-line screening system organized at the Kazusa DNA Research Institute. The principles of the screening method were as described²⁵. Gene-specific primers were 5'-ATATGCGATAGCGACTCTCGTACAA-3' and 5'-AAC-CAAAATGCATATCAATCAGCAG-3'. T-DNA (pPCV1En4HPT)²⁶ specific primers were 5'-ATAACGCTGCGGACATCTAC-3' and 5'-ATCTAGGCTTTGATAGTCAC-3'. We used four combinations of primer sets, each consisting of a gene specific primer and a T-DNA-specific primer. The position of the T-DNA insert was determined by sequencing the PCR products carrying the T-DNA-genome junctions.

Yeast experiments

The entire coding region of the CRE1a cDNA was PCR-amplified and cloned into the yeast expression vector p415CYC²⁷ under the CYC1 promoter at the *Sma*I site, generating p415CYC-CRE1. We used the QuickChange site-directed mutagenesis kit (Stratagene) to generate p415CYC-CRE1(G467D), p415CYC-CRE1(H459Q) and p415CYC-CRE1(D973N). After sequence confirmation, plasmids were introduced into *sln1Δ* (strain TM182¹⁰) or *ypd1Δ* (strain SW100¹⁷). Suspensions of transformants were spotted (about 800 cells per spot) onto drop-out media with 10 μM plant hormones as indicated in Fig. 4, with either 2% glucose or 2% galactose.

Received 16 October; accepted 8 December 2000.

1. Skoog, F. & Miller, C. O. Chemical regulation of growth and organ formation in plant tissues cultured in vitro. *Symp. Soc. Exp. Biol.* **11**, 118–131 (1957).
2. Mok, M. C. in *Cytokinins* (eds Mok, D. W. S. & Mok, M. C.) 155–166 (CRC, Boca Raton, 1994).
3. Kakimoto, T. CK11, a histidine kinase homolog implicated in cytokinin signal transduction. *Science* **274**, 982–985 (1996).
4. Sakakibara, H. *et al.* A response-regulator homologue possibly involved in nitrogen signal transduction mediated by cytokinin in maize. *Plant J.* **14**, 337–344 (1998).
5. Brandstatter, I. & Kieber, J. J. Two genes with similarity to bacterial response regulators are rapidly and specifically induced by cytokinin in *Arabidopsis*. *Plant Cell* **10**, 1009–1019 (1998).
6. Mizuno, T. His-Asp phosphotransfer signal transduction. *J. Biochem.* **123**, 555–563 (1998).
7. D'Agostino, I. B. & Kieber, J. J. Phosphorelay signal transduction: the emerging family of plant response regulators. *Trends Biochem. Sci.* **24**, 452–456 (1999).
8. Chang, C., Kwok, S. F., Bleecker, A. B. & Meyerowitz, E. M. *Arabidopsis* ethylene-response gene *ETR1*: similarity of product to two-component regulators. *Science* **262**, 539–544 (1993).
9. Schaller, G. E. & Bleecker, A. B. Ethylene-binding sites generated in yeast expressing the *Arabidopsis* *ETR1* gene. *Science* **270**, 1809–1811 (1995).
10. Maeda, T., Wurgler-Murphy, S. M. & Saito, H. A two-component system that regulates an osmosensing MAP kinase cascade in yeast. *Nature* **19**, 242–245 (1994).
11. Cary, A. J., Liu, W. & Howell, S. H. Cytokinin action is coupled to ethylene in its effects on the inhibition of root and hypocotyl elongation in *Arabidopsis thaliana* seedlings. *Plant Physiol.* **107**, 1075–1082 (1995).
12. Bleecker, A. B., Estelle, M. A. & Somerville, C. Insensitivity to ethylene conferred by a dominant mutation in *Arabidopsis thaliana*. *Science* **241**, 1086–1089 (1988).
13. Lincoln, C., Britton, J. H. & Estelle, M. Growth and development of the *axr1* mutants of *Arabidopsis*. *Plant Cell* **2**, 1071–1080 (1990).
14. Koornneef, M., Reuling, G. & Karssen, C. M. The isolation and characterization of abscisic acid-insensitive mutants of *Arabidopsis thaliana*. *Physiol. Plantar* **61**, 377–383 (1984).
15. Mähönen, A. P. *et al.* A novel two-component hybrid molecule regulates vascular morphogenesis of the *Arabidopsis* root. *Genes Dev.* **14**, 2938–2943 (2000).
16. Ueguchi, C., Koizumi, H., Suzuki, T. & Mizuno, T. Novel family of sensor histidine kinase genes in *Arabidopsis thaliana*. *Plant Cell Physiol.* (in the press).
17. Posas, F. *et al.* Yeast HOG1 MAP kinase cascade is regulated by a multistep phosphorelay mechanism in the SLN1-YPD1-SSK1 "two-component" osmosensor. *Cell* **86**, 865–875 (1996).
18. Schmitz, R. Y., Skoog, F., Playfair, A. J. & Leonard, N. J. Cytokinins: Synthesis and biological activity of geometric and position isomers of zeatin. *Plant Physiol.* **50**, 702–705 (1972).
19. Urao, T. *et al.* A transmembrane hybrid-type histidine kinase in *Arabidopsis* functions as an osmosensor. *Plant Cell* **11**, 1743–1754 (1999).
20. Miyata, S., Urao, T., Yamaguchi-Shinozaki, K. & Shinozaki, K. Characterization of genes for two-component phosphorelay mediators with a single HPT domain in *Arabidopsis thaliana*. *FEBS Lett.* **437**, 11–14 (1998).
21. Suzuki, T., Imamura, A., Ueguchi, C. & Mizuno, T. Histidine-containing phosphotransfer (HPT) signal transducers implicated in His-to-Asp phosphorelay in *Arabidopsis*. *Plant Cell Physiol.* **39**, 1258–1268 (1998).
22. Kubo, M. & Kakimoto, T. The CYTOKININ-HYPERSENSITIVE genes of *Arabidopsis* negatively regulate the cytokinin-signaling pathway for cell division and chloroplast development. *Plant J.* **23**, 385–394 (2000).
23. Becker, D., Kemper, E., Schell, J. & Masterson, R. New binary vectors with selectable markers located proximal to the left T-DNA border. *Plant Mol. Biol.* **20**, 1195–1197 (1992).
24. Akama, K. *et al.* Efficient transformation of *Arabidopsis thaliana*: comparison of the efficiencies with various organs, plant ecotypes and *Agrobacterium* strains. *Plant Cell Rep.* **12**, 7–11 (1992).
25. McKinney, E. C. *et al.* Sequence-based identification of T-DNA insertion mutations in *Arabidopsis*: actin mutants *act2-1* and *act4-1*. *Plant J.* **8**, 613–622 (1995).
26. Hayashi, H., Czaja, I., Lubenow, H., Schell, J. & Walden, R. Activation of a plant gene by T-DNA tagging: auxin-independent growth in vitro. *Science* **258**, 1350–1353 (1992).
27. Mumberg, D., Müller, R. & Funk, M. Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. *Gene* **156**, 119–122 (1995).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

M. Higuchi carried out most of the yeast work. We thank H. Saito for yeast strains,

E. Kemper for pGPTV-KAN and K. Torii for comments. Seeds of *abi3-1*, *aux1-7* and *ein2-1* were obtained from ABRC. p415CYC was obtained from ATCC. This study was in part supported by grants from the Ministry of Education, Science and Culture of Japan, and from the Science and Technology Agency to T.K.

Correspondence and requests for materials should be addressed to T.K. (e-mail: kakimoto@bio.sci.osaka-u.ac.jp). The accession numbers for CRE1a and CRE1b are AB049934 and AB049935, respectively.

cdc2 links the *Drosophila* cell cycle and asymmetric division machineries

Murni Tio, Gerald Udolph*, Xiaohang Yang and William Chia

Institute of Molecular and Cell Biology, 30 Medical Drive, 117609 Singapore

Asymmetric cell divisions can be mediated by the preferential segregation of cell-fate determinants into one of two sibling daughters. In *Drosophila* neural progenitors, Inscuteable^{1–3}, Partner of Inscuteable^{4,5} and Bazooka^{6,7} localize as an apical cortical complex at interphase, which directs the apical–basal orientation of the mitotic spindle as well as the basal/cortical localization of the cell-fate determinants Numb^{8,9} and/or Prospero^{10,11} during mitosis. Although localization of these proteins shows dependence on the cell cycle, the involvement of cell-cycle components in asymmetric divisions has not been demonstrated. Here we show that neural progenitor asymmetric divisions require the cell-cycle regulator *cdc2*. By attenuating *Drosophila* *cdc2* function without blocking mitosis, normally asymmetric progenitor divisions become defective, failing to correctly localize asymmetric components during mitosis and/or to resolve distinct sibling fates. *cdc2* is not necessary for initiating apical complex formation during interphase; however, maintaining the asymmetric localization of the apical components during mitosis requires Cdc2/B-type cyclin complexes. Our findings link *cdc2* with asymmetric divisions, and explain why the asymmetric localization of molecules like Inscuteable show cell-cycle dependence.

The embryonic central nervous system (CNS) of *Drosophila* is derived from progenitors called neuroblasts (NBs). NBs undergo repeated asymmetric divisions, budding off a series of ganglion mother cells (GMCs) from their basal/lateral surfaces; GMCs can divide asymmetrically to produce progeny with distinct neuronal fates. Both the NB and GMC asymmetric divisions are mediated, in part, by a protein localization machinery that directs the preferential segregation of Prospero (Pros) or Numb to the more basally located daughter. Mitosis is driven by activation of the Cdc2 protein kinase, which, during the first 13 embryonic divisions, depends on dephosphorylation by the product of maternal *string* (*cdc25*)^{12,13}. NB divisions occur after depletion of maternal *string* and depend on zygotic *string*. However, NBs, although arrested at G2 of cycle 14, do form in embryos lacking zygotic *string*. In contrast, loss of zygotic *cdc2* does not substantially affect embryonic development, and lethality occurs during postembryonic development^{14,15}.

From a mutant screen we identified two lines that exhibited defective localization of Pros and Inscuteable (Insc) in NBs. Genetic mapping, complementation and DNA sequencing revealed that the phenotypes associated with both mutants were caused by the same mutation in *cdc2*, resulting in a glutamic acid to glutamine change at amino-acid 51. Embryos homozygous for *cdc2*^{E51Q} show late

* Present address: MRC Centre for Developmental Neurobiology, King's College London, London SE1 1UL, UK.

embryonic lethality and exhibit various abnormalities that are also seen in *insc* and *partner of inscuteable* (*pins*) mutants, which can be explained by defects in asymmetric divisions.

To illustrate these defects, we focused on the first GMC produced from NB4-2, GMC4-2a, which divides to generate two daughter neurons, RP2 and its sibling RP2sib¹⁶. The Even-skipped (*Eve*) protein is expressed in the GMC4-2a sublineages (Fig. 1a). *Eve* is initially expressed in both RP2 and RP2sib; however, its expression is extinguished in RP2sib, such that late in embryonic development only one *Eve*⁺ neuron can be seen at the RP2 position in each wild-type hemineuromere. Two types of defects are seen in *cdc2*^{E51Q} homozygotes. Fourteen per cent of the mutant hemisegments exhibit near the RP2 position a single *Eve*⁺ cell that has all of the characteristics of the RP2 neuron but is larger than the wild-type RP2 neuron. In cell-division mutants that prevent GMC4-2a division, GMC4-2a differentiates into an RP2-like cell that is larger than normal^{17,18} (Fig. 1b; Table 1). More notably, 33% of the mutant hemisegments possess two cells near the RP2 position, both of which show characteristics of the RP2 neuron as judged by marker expression (Fig. 1b, d; Table 1).

Duplication of the RP2 neuron is caused by an RP2sib → RP2 transformation, indicating that the normally asymmetric division of GMC4-2a (GMC4-2a → RP2 + RP2sib), has been converted to a symmetric division (GMC4-2a → RP2 + RP2) in the mutant. This defect in asymmetric division is not restricted to the CNS; the normally asymmetric division of muscle progenitor, P15, which in wild type produces a single *Eve*-expressing muscle DA1 and a second daughter of unknown fate (Fig. 1e), can also become symmetric in mutant embryos, leading to the duplication of muscle DA1¹⁹ (Fig. 1f).

We assessed the localization of *Insc*, *Partner of Numb*²⁰ (*Pon*, which colocalizes with *Numb*) and *Miranda*^{21–23} (which colocalizes with *Pros*) in mutant neural progenitors which are clearly under-

going mitosis. Dividing wild-type GMC4-2a always localizes *Insc* as an apical crescent (20/20; Fig. 2a) and *Pon* as a basal crescent (20/20; Fig. 2d). In dividing *cdc2*^{E51Q} GMC4-2a, defective localization of *Insc* (39%; 9/23; Fig. 2b, c) and *Pon* (37%; 7/19; Fig. 2e, f), in the form of cortical distribution or misplaced crescents, and mis-orientation of the mitotic spindle (Fig. 2f) are observed.

These defects are not restricted to GMCs. Mislocalization of *Insc* (25%, *n* = 246; Fig. 2h), *Bazooka* (38%, *n* = 85; data not shown), *Pon* (Fig. 2k) and *Miranda* (data not shown), and defective spindle orientation (Fig. 2n, o) are also seen in mutant mitotic NBs. These data suggest that the underlying cause of the abnormal progenitor divisions may be the failure to localize the apical components during mitosis, and consequently localization of the basal determinants is also defective. Consistent with this notion, the duplicated RP2 neurons show identical nuclear size—a phenotype characteristic of *insc* and *pins* mutants^{4,17}.

Embryonic lethality and defects in asymmetric division are seen in *cdc2*^{E51Q} embryos but embryos totally lacking zygotic *cdc2* function develop essentially normally (Table 1), owing to the maternal contribution of *cdc2*. Several observations indicate that *cdc2*^{E51Q} acts as a maternal effect dominant-negative allele that can antagonize the maternally inherited wild-type *cdc2*. Strong mutant phenotypes are seen in genotypically hemizygous (*cdc2*^{E51Q}/deficiency) embryos only if the *cdc2*^{E51Q} allele is inherited from the (*cdc2*^{E51Q}/CyO) mother, and not if it comes from the father (Table 1, (2)). Moreover an earlier arrest of cell divisions, resulting in a marked increase in the frequency of undivided GMC4-2a cells, is seen by overexpressing *cdc2*^{E51Q} (from a *uas-cdc2*^{E51Q} transgene) (Table 1, (3)).

To show that defects in asymmetric cell divisions and protein localization are not peculiar to *cdc2*^{E51Q}, we used a stock that is homozygous for an amorphic allele, *cdc2*^{B47}, and in addition contains four copies of a transgene carrying a temperature-sensitive

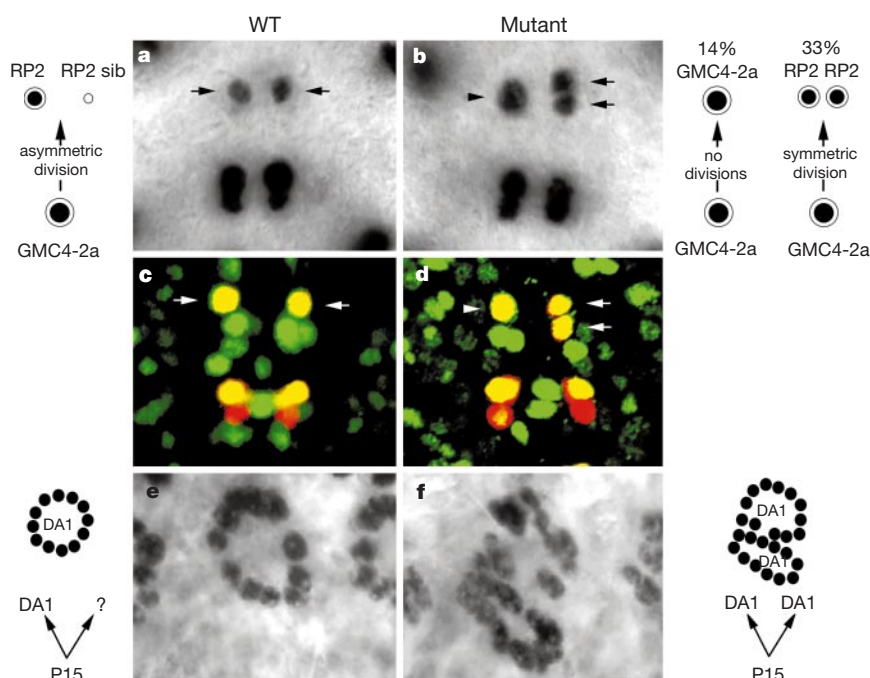


Figure 1 Conversion of asymmetric to symmetric divisions in *cdc2* mutants. **a, b**, Wild-type GMC4-2a divides to produce *Eve*⁺ neurons, RP2 and RP2sib; only RP2 retains *Eve* expression, and by stage 14 only a single *Eve*⁺ neuron can be seen at the RP2 position (arrow, **a**). In *cdc2*^{E51Q} embryos, 14% of GMC4-2a fail to divide and differentiate into a 'large' RP2 (arrowhead, **b**); 33% of GMC4-2a divide symmetrically to give rise to two RP2s with equal nuclear size (arrows, **b**). **c, d**, In addition to *Eve* (red) both wild-type RP2 (**c**) and

the duplicated RP2 and 'large' RP2 found in the mutant (**d**) express *Zfh-1* (green). **e, f**, Muscle progenitor P15 divides to produce one *Eve*⁺ founder for muscle DA1 and a sibling of unknown fate (schematically represented); the single muscle DA1 with about 10 *Eve*⁺ nuclei can be seen from a dorsal-lateral view of a stage 14/15 wild-type embryo (**e**). In *cdc2*^{E51Q} (not shown) or *cdc2ts4X* embryos upshifted to the non-permissive temperature, duplication of the DA1 muscle can be seen (**f**).

allele, *cdc2*^{A171T}, of *cdc2* (referred to as *cdc2ts4X*)²⁴. Immunoprecipitates from *cdc2ts4X* embryonic extracts exhibit kinase activity that is highly temperature sensitive²⁴. Under appropriate temperature-shift conditions, these embryos can exhibit all of the phenotypes seen in *cdc2*^{E51Q} including RP2 (Table 1, (1)) and muscle DA1 (Fig. 1f) duplications, and defective protein localization in dividing progenitors (Fig. 2i, l; and data not shown).

These results suggest that the levels of Cdc2 activity determine whether and how a progenitor divides. When the level of Cdc2 is low, neural progenitors fail to divide; at intermediate levels they can divide, but mitotic NBs often fail to correctly localize asymmetric components, and GMC divisions can become symmetric with respect to segregation of cell-fate determinants and the size and fate of their daughters; only at higher levels of *cdc2* do normal asymmetric divisions take place.

We used a binary expression system²⁵ to express wild-type and mutated forms of *cdc2* in the neural progenitors of *cdc2*^{E51Q} embryos. Expression of a wild-type *cdc2* transgene rescued the *cdc2*^{E51Q} phenotypes, whereas the expression of enzymatically dead versions²⁶ of *cdc2*, *cdc2*^{T161A} and *cdc2*^{K33R/T161A}, which do not exhibit dominant-negative properties, did not rescue those phenotypes (Table 1, (4)), suggesting that kinase activity is required for

asymmetric divisions. As Cdc2 kinase activity appears to be required to maintain apical Insc localization during mitosis, might it also be required for the apical localization of Insc during interphase? In embryos lacking zygotic *string*, NBs form but arrest at G2 during interphase of cell-cycle stage 14 and fail to enter mitosis because mitotic kinase activation does not occur. Insc (Fig. 3a), Pins and Bazooka (data not shown) form normal apical crescents, indicating that the initial apical localization of the apical components during interphase does not require mitotic kinase activation. Moreover, temperature upshifts can induce similar *string*-like phenotypes in *cdc2ts4X* embryos, and in these embryos the NBs arrested at G2 show apical localization of the apical components (Fig. 3b; and data not shown). These data suggest that Cdc2 kinase activity is required to maintain the apical complex proteins during

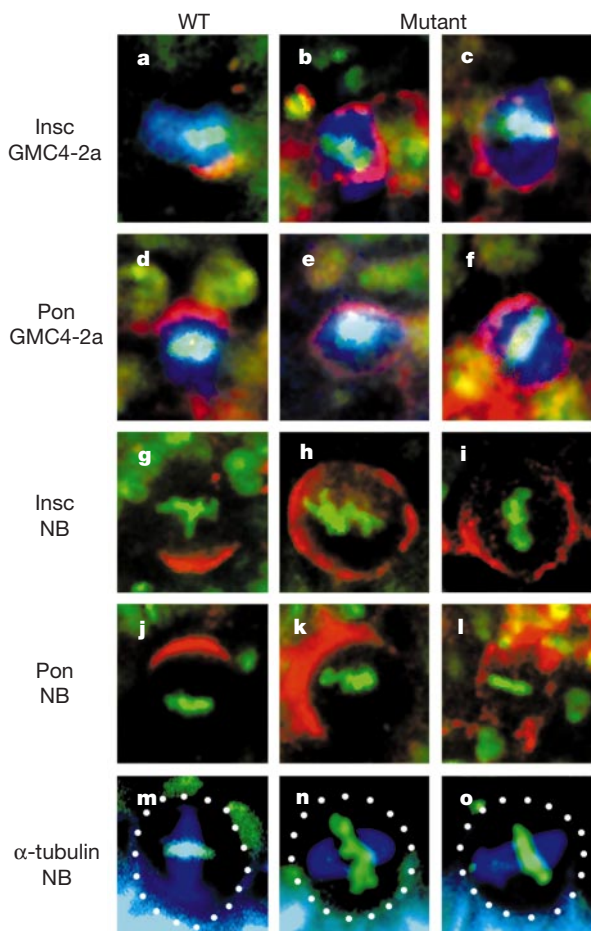


Figure 2 Defective protein localization in dividing neural progenitors of *cdc2* mutants. Lateral optical sections of stage10/11 embryos. Apical is down. **a–c**, Wild-type (**a**) and *cdc2*^{E51Q} (**b, c**) mitotic GMC4-2a labelled with anti-Eve (blue), DNA stain (green) and anti-Insc (red). **d–f**, Wild-type (**d**) and *cdc2*^{E51Q} (**e, f**) mitotic GMC4-2a labelled with anti-Eve (blue), DNA stain (green) and anti-Pon (red). **g–i**, Wild-type (**g, h**), *cdc2*^{E51Q} (**h, k**) and *cdc2ts4X* (**i, l**) mitotic NBs labelled with DNA stain (green) and anti-Insc (red, **g–i**) or anti-Pon (red, **j–l**). **m–o**, Wild-type (**m**) and *cdc2*^{E51Q} (**n, o**) mitotic NBs labelled with DNA stain (green) and anti-α-tubulin (blue). Note the mislocalization of Insc and Pon and the misorientation of mitotic spindles in mutant GMCs and NBs.

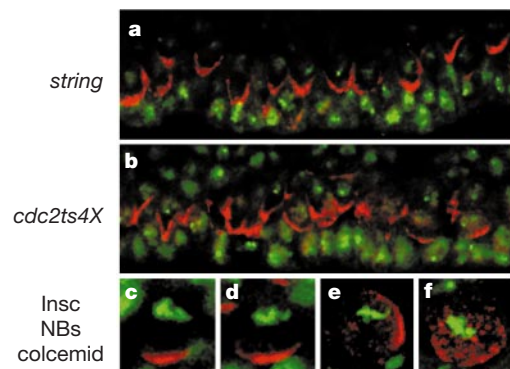


Figure 3 Cdc2 kinase activity is required for the maintenance but not the establishment of Insc apical localization. A *string* embryo arrested at interphase of cycle14 (**a**) and a *cdc2ts4X* embryo shifted to the non-permissive temperature arrested at interphase of cycle 14 (**b**) stained with anti-Insc (red) and DNA stain (green). Note the apical Insc localization in the NBs of both embryos. **c–f**, Wild-type NB treated with colcemid and shifted to 31 °C (**c**); *cdc2ts4X* NB treated with colcemid and maintained at 21 °C (**d**); *cdc2ts4X* NB treated with colcemid and shifted to 31 °C (**e, f**). Insc (red) and DNA (green). Apical is down.

Table 1 Phenotypic analysis of various *cdc2* mutations and genotypes

	RP2 duplication	Big RP2s	n	Lethal phase
Wild type	0%	0%	200	
<i>cdc2</i> ^{E51Q} (amorph)	33%	14%	266	Embryonic
<i>cdc2</i> ^{D57} (amorph)	0.8%	0.8%	381	Larval/pupal
<i>cdc2</i> ^{E1-23} (amorph)	0.25%	0.8%	390	Larval/pupal
<i>cdc2</i> ^{B47} (amorph)	3.6%	3%	307	Larval/pupal
(1) <i>cdc2ts4X</i> *	42.7%	4.6%	144	
(2) <i>cdc2</i> ^{E51Q} /Df(2L)J3 from <i>cdc2</i> ^{E51Q} /Cyto(fz-lacZ) mothers	33%	7%	249	
(2) <i>cdc2</i> ^{E51Q} /Df(2L)J3 from <i>cdc2</i> ^{E51Q} /Cyto(fz-lacZ) fathers	2%	1%	565	
(3) <i>cdc2</i> ^{E51Q} <i>sca-Gal4</i> (1X); <i>uas-cdc2</i> ^{E51Q} (1X)	15%	52%	690	
(4) <i>cdc2</i> ^{E51Q} <i>uas-wt cdc2</i> (1X); <i>da-Gal4</i> (1X)	0%	0%	486	
(4) <i>cdc2</i> ^{E51Q} <i>uas-T161A</i> (1X); <i>da-Gal4</i> (1X)	25.5%	10.5%	475	
(4) <i>cdc2</i> ^{E51Q} <i>uas-K33R, T161A</i> (1X); <i>da-Gal4</i> (1X)	22.4%	12%	759	

n is the number of hemisegments scored. The *cdc2* alleles specified represent homozygotes unless otherwise indicated. (1) A viable, fertile stock referred to as *cdc2ts4X*²⁴ (see text). (2) Genotypically *cdc2*^{E51Q}/Df(2L)J3 embryos show different phenotypes depending on whether the *cdc2*^{E51Q} allele is maternally or paternally inherited. (3) Overexpression of a *cdc2*^{E51Q} transgene using a pan-neural driver, *sca-Gal4*, has dominant-negative effects, increasing the proportion of GMC4-2a which fails to divide. (4) Experiments in which wild-type and enzymatically compromised *cdc2* transgenes were overexpressed in a *cdc2*^{E51Q} homozygous background (using a ubiquitous driver, *da-Gal4*), showing the requirement of kinase activity in asymmetric cell-fate specification. The T161A and the K33R mutations would be expected to compromise cyclin binding and ATP hydrolysis, respectively. Mutant *cdc2* genes were generated by PCR and authenticated by DNA sequencing; germline transformants were produced using standard methods. Temperature shifts were performed as described in Methods.

* The frequency of RP2 duplications is extremely variable in upshifted *cdc2ts4X* embryos and our quantitation was based on ten embryos with the highest frequency of duplications. This phenotype was never seen in wild-type embryos processed in parallel.

mitosis but not for their establishment during interphase.

If Cdc2 activity is responsible for maintaining the apical components, premature inactivation/reduction of its activity—at a point in mitosis when its activity is normally high—should lead to premature delocalization of apical components like Insc. *cdc2ts4X* and wild-type control embryos were arrested at metaphase using colcemid treatment at 21 °C for 30 min (see Methods); half of the embryos were shifted to 31 °C for 45 min, while the other half were kept at 21 °C for 45 min, then both groups were fixed and stained for Insc. *cdc2ts4X* NBs arrested at metaphase and maintained at 21 °C showed normal apical crescents of Insc (98%, $n = 270$; Fig. 3d); similarly, colcemid-treated wild-type controls that were shifted to 31 °C show 100% apical localization ($n = 106$; Fig. 3c). *cdc2ts4X* NBs arrested at metaphase and upshifted to 31 °C showed defective localization of Insc (only 7% have normal apical crescents; $n = 82$; Fig. 3e, f) and Miranda (data not shown). These results show that downregulating Cdc2 activity in NBs

arrested at metaphase can cause delocalization of the apical (and basal) component proteins and provide direct evidence that elevated Cdc2 activity is required to maintain the localization of apical components during mitosis.

There are three known mitotic cyclins in *Drosophila*, cyclin A, B and B3; these need to be destroyed for mitotic exit to occur²⁷. Is a subset of these cyclins preferentially required for maintaining the apical components? We followed the temporal profile of Insc localization with respect to the time of cyclin A (metaphase), cyclin B (early anaphase) and cyclin B3 (late anaphase) destruction. The results from anti-Insc/anti-cyclin-A/DNA-stain triple-labelling experiments (Fig. 4b) show that Insc remains apically localized at metaphase/anaphase after destruction of cyclin A. Supporting the idea that cyclin A is dispensable, we failed to detect any evidence of defective Insc localization in cyclin A single and cyclin A, cyclin B3 or cyclin A, cyclin B double mutants (data not shown). In wild-type NBs, Insc delocalization occurs after chromosome separation, coinciding with the time when cyclin B3 becomes undetectable (Fig. 4b). Single mutants in cyclin B and cyclin B3 do not show defects in Insc localization (data not shown); however, in cyclin B, cyclin B3 double mutants, mislocalization of Insc can be seen in most mitotic NBs (72%, $n = 120$ for prophase; 71%, $n = 90$ for metaphase; Fig. 4c). These results indicate that whereas cyclin A appears dispensable, the B-type cyclins are required to maintain asymmetric localization of Insc during mitosis.

Our data show that a key cell-cycle regulator is involved in mediating asymmetric cell divisions. Phosphorylation mediated by Cdc2 is likely to be important in maintaining the correct localization of the apical complex of asymmetry proteins during mitosis; however, Cdc2 probably does not act directly on Insc as the putative Cdc2 phosphorylation sites of Insc can be removed without affecting its function in overexpression paradigms^{28,29}. These observations provide an explanation for the normal temporal profile of the localization of molecules like Insc. As long as there is Cdc2 kinase activity during mitosis, apical/cortical localization of molecules such as Insc is maintained; however, when kinase (and B-type cyclins) is destroyed towards the end of mitosis, the apical components become delocalized. Our findings indicate that the tight temporal correlation of asymmetrically localized components important for mediating asymmetric divisions to the cell cycle may be because the two processes share key regulator(s) like *cdc2*. □

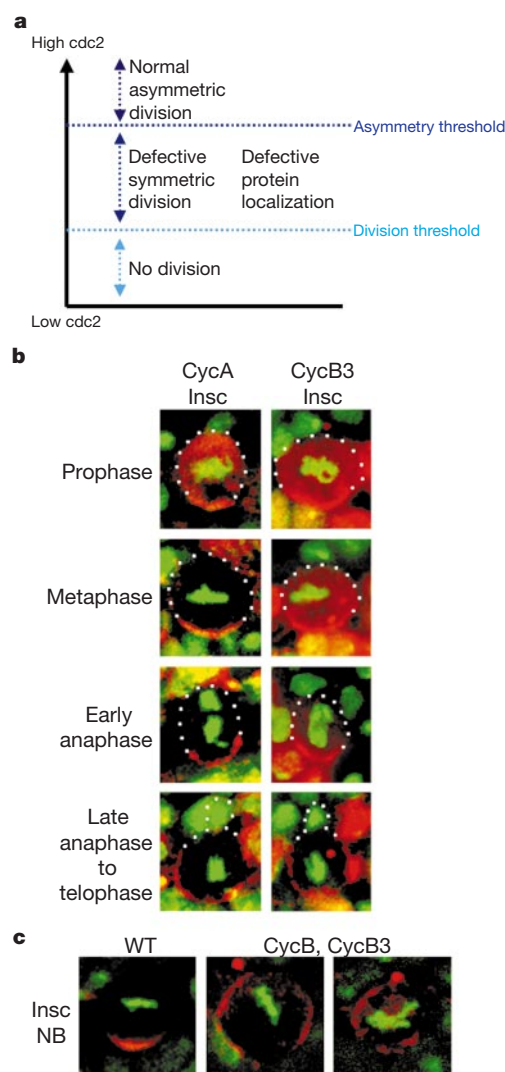


Figure 4 Cdc2 kinase activity is required for asymmetric cell divisions. **a**, Model proposing that different thresholds of kinase activity determine whether and how a progenitor divides (see text). **b**, Wild-type mitotic NBs are stained with anti-Insc (red, apical/cortical), DNA-stain (green) and anti-cyclin-A or anti-cyclin-B3 (red, cytoplasmic). Note that Insc remains localized at metaphase and early anaphase when cyclin A cannot be detected but cyclin B3 levels remain significant. When Insc becomes delocalized (expanded) at late anaphase/telophase, cyclin B3 cannot be detected. **c**, Wild-type metaphase NB (left) and cyclin B, cyclin B3 double-mutant metaphase (middle) and prophase (right) NBs; note the premature delocalization of Insc (red) in the double-mutant NBs.

Methods

The initial mutant screen

A collection of ~300 lethal enhancer trap insertions that show β -galactosidase expression in all or subsets of NBs were identified and collected over the past few years. We screened embryos from these insertions lines using confocal microscopy after staining with anti-Insc and anti-Pros. The E51Q allele associated with the two lines showing defective Pros and Insc localization were unconnected with their respective P-element insertions, and most probably originated from a pre-existing mutation in the progenitor chromosome used to generate the enhancer traps.

Flies and embryos

The following *cdc2* alleles were used: *cdc2^{E51Q}*, *cdc2ts4X*, *cdc2^{D57}*, *cdc2^{E1-23}* and *cdc2^{B47}*. The following cyclin alleles were used: *cyclinA^{C8LR1}*, *cyclinB²*, *cyclinB3³*, *cyclinB3³*. *string/TM3 Ubx-lacZ* was used. All lethal mutant alleles that we used in this study were balanced over balancer chromosomes containing P-element insertions expressing *ftz-lacZ* or *Ubx-lacZ*, and homozygous mutant embryos were identified by the absence of β -galactosidase expression. *cdc2ts4X* was raised at 18 °C, and all other stocks were raised at 25 °C. For the temperature-shift experiments designed to examine neuronal or muscle cell fate changes, *cdc2ts4X* embryos were collected for 5 h and aged for 7 h at 21 °C, shifted to 31 °C for 1 h, and aged at 21 °C overnight until fixation. For the temperature-shift experiments designed to examine protein localization in NBs and GMCs, *cdc2ts4X* embryos were dechorionated and incubated in a 31 °C water bath for 1 h before fixation. Wild-type control embryos were processed in parallel.

Drug treatment

Embryos were permeabilized in octane¹² and treated with 20 μ g ml⁻¹ colcemid for 30 min

at 21 °C. They were either maintained at 21 °C or upshifted to 31 °C for a further 45 min before fixation and staining.

Immunohistochemistry

We performed fixation and antibody staining of embryos as described^{4,28}. We used the following primary antibodies: mouse anti- α -tubulin, mouse anti-cyclin-A, mouse anti-cyclin-B, rabbit anti-cyclin-B3, rabbit and mouse anti-Eve, mouse anti-Zfh-1, rabbit and mouse anti- β -galactosidase, rabbit anti-Insc, rabbit anti-Miranda, rabbit anti-Pon, mouse anti-Pros. DNA was visualized with a DNA fluorochrome³⁰. Confocal microscopy was performed using a MRC1024 scanhead.

Received 1 November; accepted 7 December 2000.

- Kraut, R. & Campos-Ortega, J. A. *inscuteable*, a neural precursor gene of *Drosophila*, encodes a candidate for a cytoskeleton adaptor protein. *Dev. Biol.* **174**, 65–81 (1996).
- Kraut, R., Chia, W., Jan, L. Y., Jan, Y. N. & Knoblich, J. A. Role of *inscuteable* in orienting asymmetric cell divisions in *Drosophila*. *Nature* **383**, 50–55 (1996).
- Li, P., Yang, X., Wasser, M., Cai, Y. & Chia, W. Inscuteable and Staufin mediate asymmetric localization and segregation of *prospero* RNA during *Drosophila* neuroblast cell divisions. *Cell* **90**, 437–447 (1997).
- Yu, F., Morin, X., Cai, Y., Yang, X. & Chia, W. Analysis of *partner of inscuteable*, a novel player of *Drosophila* asymmetric divisions, reveals two distinct steps in Inscuteable apical localization. *Cell* **100**, 399–409 (2000).
- Schaefer, M., Shevchenko, A., Shevchenko, A. & Knoblich, J. A. A protein complex containing Inscuteable and the G α -binding protein Pins orients asymmetric cell divisions in *Drosophila*. *Curr. Biol.* **10**, 353–362 (2000).
- Wodarz, A., Ramrath, A., Kuchinke, U. & Knust, E. Bazooka provides an apical cue for Inscuteable localization in *Drosophila* neuroblasts. *Nature* **402**, 544–547 (1999).
- Schober, M., Schaefer, M. & Knoblich, J. Bazooka recruits Inscuteable to orient asymmetric cell divisions in *Drosophila* neuroblasts. *Nature* **402**, 548–551 (1999).
- Rhyu, M. S., Jan, L. Y. & Jan, Y. N. Asymmetric distribution of Numb protein during division of the sensory organ precursor cell confers distinct fates to daughter cells. *Cell* **76**, 477–491 (1994).
- Knoblich, J. A., Jan, L. Y. & Jan, Y. N. Asymmetric segregation of Numb and Prospero during cell division. *Nature* **377**, 624–627 (1995).
- Hirata, J., Nakagoshi, H., Nabeshima, Y. & Matsuzaki, F. Asymmetric segregation of the homeo-domain protein Prospero during *Drosophila* development. *Nature* **377**, 627–630 (1995).
- Spana, E. & Doe, C. Q. The Prospero transcription factor is asymmetrically localised to the cell cortex during neuroblast mitosis in *Drosophila*. *Development* **121**, 3187–3195 (1995).
- Edgar, B. A. & O'Farrell, P. H. The three postblastoderm cell cycles of *Drosophila* embryogenesis are regulated in G2 by *string*. *Cell* **62**, 469–480 (1990).
- Edgar, B. A., Sprenger, F., Duronio, R. J., Leopold, P. & O'Farrell, P. H. Distinct molecular mechanisms regulate cell cycle timing at successive stages of *Drosophila* embryogenesis. *Genes Dev.* **8**, 440–452 (1994).
- Clegg, N. J., Whitehead, I. P., Williams, J. A., Spiegelman, G. B. & Grigliatti, T. A. A developmental and molecular analysis of *cdc2* mutations in *Drosophila melanogaster*. *Genome* **36**, 676–685 (1993).
- Stern, B., Ried, G., Clegg, N. J., Grigliatti, T. A. & Lehner, C. F. Genetic analysis of the *Drosophila cdc2* homolog. *Development* **117**, 219–232 (1993).
- Doe, C. Q. Molecular markers for identified neuroblasts and ganglion mother cells in the *Drosophila* central nervous system. *Development* **116**, 855–863 (1992).
- Buescher, M. *et al.* Binary sibling neuronal cell fate decisions in the *Drosophila* embryonic central nervous system are nonstochastic and require *inscuteable*-mediated asymmetry of ganglion mother cells. *Genes Dev.* **12**, 1858–1870 (1998).
- Lear, B. C., Skeath, J. B. & Patel, N. H. Neural cell fate in *rcal* and *cycA* mutants: the roles of intrinsic and extrinsic factors in asymmetric division in the *Drosophila* central nervous system. *Mech. Dev.* **88**, 207–219 (1999).
- Carmena, A., Murugasu-Oei, B., Menon, D., Jimenez, F. & Chia, W. *inscuteable* and *numb* mediate asymmetric muscle progenitor cell divisions during *Drosophila* myogenesis. *Genes Dev.* **12**, 304–315 (1998).
- Lu, B., Rothenberg, M., Jan, L. Y. & Jan, Y. N. Partner of Numb colocalises with Numb during mitosis and directs Numb asymmetric localisation in *Drosophila* neural and muscle progenitors. *Cell* **95**, 225–235 (1998).
- Shen, C. P., Jan, L. Y. & Jan, Y. N. Miranda is required for the asymmetric localisation of Prospero during mitosis in *Drosophila*. *Cell* **90**, 449–458 (1997).
- Ikeshima-Kataoka, H., Skeath, J. B., Nabeshima, Y., Doe, C. Q. & Matsuzaki, F. Miranda directs Prospero to a daughter cell during *Drosophila* asymmetric divisions. *Nature* **390**, 625–629 (1997).
- Schuldt, A. J. *et al.* Miranda mediates asymmetric protein and RNA localisation in the developing nervous system. *Genes Dev.* **12**, 1847–1857 (1998).
- Sigrist, S., Ried, G. & Lehner, C. F. Dmcd2 kinase is required for both meiotic divisions during *Drosophila* spermatogenesis and is activated by the Twine/Cdc25 phosphatase. *Mech. Dev.* **53**, 247–260 (1995).
- Brand, A. H. & Perrimon, N. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development* **118**, 401–415 (1993).
- Ducommun, B. *et al.* *cdc2* phosphorylation is required for its interaction with cyclin. *EMBO J.* **10**, 3311–3319 (1991).
- Sigrist, S., Jacobs, H., Stratmann, R. & Lehner, C. F. Exit from mitosis is regulated by *Drosophila fizzy* and the sequential destruction of cyclins A, B and B3. *EMBO J.* **14**, 4827–4838 (1995).
- Tio, M., Zavortink, M., Yang, X. & Chia, W. A functional analysis of *inscuteable* and its roles during *Drosophila* asymmetric cell divisions. *J. Cell Sci.* **112**, 1541–1551 (1999).
- Knoblich, J. A., Jan, L. Y. & Jan, Y. N. Deletion analysis of the *Drosophila* Inscuteable protein reveals domains for cortical localization and asymmetric localization. *Curr. Biol.* **9**, 155–158 (1999).
- Lundell, M. J. & Hirsh, J. A new visible light DNA fluorochrome for confocal microscopy. *Biotechniques* **16**, 434–440 (1994).

Acknowledgements

We thank C. Q. Doe, M. Frasch, D. Glover, M. Harrington, Y.-N. Jan, E. Knust, Z. Lai, K. Matthews, F. Matsuzaki, P. O'Farrell, R. Saint, K. Zinn, the Developmental Studies Hybridoma Bank, Bloomington and Umea Stock Centres, and C. Lehner in particular for stocks and antibodies; F. S. Hing and C. T. Ong for technical support; M. Zavortink, U. Surana and members of our laboratory for comments and discussions; and IMCB for financial support. G.U. is supported by the Wellcome Trust.

Correspondence and requests for materials should be addressed to W.C. (e-mail: mcbwchia@imcb.nus.edu.sg).

Opposing effects of Ets and Id proteins on p16^{INK4a} expression during cellular senescence

Naoko Ohtani*, Zoe Zebedee*, Thomas J. G. Huot†, Julie A. Stinson‡, Masataka Sugimoto*§, Yasuhiro Ohashi||, Andrew D. Sharrocks‡, Gordon Peters† & Eiji Hara*¶

* CRC Cell Cycle Group, Paterson Institute for Cancer Research, Christie Hospital NHS Trust, Manchester M20 4BX, UK

† Imperial Cancer Research Fund Laboratories, London WC2A 3PX, UK

‡ School of Biological Sciences, University of Manchester, Manchester M13 9PT, UK

§ Department of Immunology, Juntendo University Medical School, Tokyo 113-8421, Japan

|| Laboratory of Molecular Biology, Nihon Schering K.K., Osaka 532-0004, Japan

¶ Department of Biomolecular Sciences, University of Manchester Institute of Science and Technology (UMIST), Manchester M60 1QD, UK

The p16^{INK4a} cyclin-dependent kinase inhibitor¹ is implicated in replicative senescence, the state of permanent growth arrest provoked by cumulative cell divisions or as a response to constitutive Ras–Raf–MEK signalling in somatic cells^{2–8}. Some contribution to senescence presumably underlies the importance of p16^{INK4a} as a tumour suppressor⁹ but the mechanisms regulating its expression in these different contexts remain unknown. Here we demonstrate a role for the Ets1 and Ets2 transcription factors¹⁰ based on their ability to activate the p16^{INK4a} promoter through an ETS-binding site and their patterns of expression during the lifespan of human diploid fibroblasts. The induction of p16^{INK4a} by Ets2, which is abundant in young human diploid fibroblasts, is potentiated by signalling through the Ras–Raf–MEK kinase cascade and inhibited by a direct interaction with the helix–loop–helix protein Id1 (ref. 11). In senescent cells, where the Ets2 levels and MEK signalling decline, the marked increase in p16^{INK4a} expression is consistent with the reciprocal reduction of Id1 and accumulation of Ets1.

The p16^{INK4a} tumour suppressor protein functions as an inhibitor of Cdk4 and Cdk6, the cyclin-dependent kinases that initiate the phosphorylation of the retinoblastoma protein, pRb^{9,12}. Thus, p16^{INK4a} has the capacity to arrest cells in the G1 phase of the cell cycle and its probable physiological role is in the implementation of an irreversible growth arrest termed replicative senescence. Senescence occurs naturally when cultured cells reach the end of their lifespan^{13,14}, triggered by the inexorable loss of telomeric DNA^{15,16}, but a similar phenotype can be induced when primary cells are challenged by an activated Ras oncogene or its downstream effectors Raf and MEK^{6–8}. In each case the arrest is accompanied by elevated levels of p16^{INK4a} (refs 2–8) and ectopic expression of p16^{INK4a} in young human diploid fibroblasts (HDFs) the senescent phenotype^{8,17}.

To delineate the signalling pathways that regulate p16^{INK4a} expression in these different contexts, we asked whether existing

p16^{INK4a} promoter constructs³ were responsive to activated Ras or MEK. The luciferase activity of a reporter plasmid containing 869 nucleotides upstream of the p16^{INK4a} translation start site (designated -869) was stimulated approximately 5-fold by H-RasV12 or constitutively active MEK in SV40-immortalized human fibroblasts (Fig. 1a). The effect was blocked by PD98059, an inhibitor of MEK, but was barely affected by Wortmannin, an inhibitor of PI-3 kinase (Fig. 1a). Truncation of the promoter reduced the basal activity, as previously reported³, but responsiveness to Ras and MEK was still apparent in the shortest construct tested (Fig. 1b). Although the p16^{INK4a} promoter sequences are not well conserved between human and mouse, both include potential binding sites for Ets-family transcription factors (Fig. 1c). Members of the Ets family, which includes Ets1, Ets2, PEA3, SAP1 and ELK1, are known to be downstream targets of Ras–Raf–MEK signalling and can be activated by MAPK-mediated phosphorylation¹⁰. Both Ets1 and Ets2 increased the activity of the -869 promoter construct by 5- to 10-fold and the effects were further potentiated by co-expression of activated MEK (Fig. 1d) or activated Ras (not shown). In contrast, PEA3, SAP1 and ELK1 caused a 2- to 3-fold reduction in p16^{INK4a} promoter activity (Fig. 1d), which suggests that they might act as negative competitors, although MEK was able to counteract the suppressive effects of SAP1. The DNA-binding domain of Ets2 (E2DBD), which acts as a dominant negative inhibitor for both Ets1 and Ets2 (ref. 18) completely negated the activation observed with Ets1 and Ets2, with or without MEK (Fig. 1d). Two mutations were introduced in the putative Ets-binding site in the human p16^{INK4a}

promoter. Mut1, in which four nucleotides in the two overlapping Ets consensus sites were changed (Fig. 1c) showed a reduced response to Ets2 (Fig. 1e). In Mut2, two further residues were changed to remove a third potential binding sequence at -117 to -122, only present in the human promoter, but there was no further loss of responsiveness to Ets2. To confirm that these findings were not specific to SVts8 cells, the promoter assays, using the -869 construct, were reproduced in the TIG3 strain of primary HDFs (Fig. 1f). Together, these results implied that Ets1 and Ets2 can activate the p16^{INK4a} promoter through the conserved Ets-binding site.

To substantiate these ideas, we sought evidence that Ets1 or Ets2 can bind to and influence the endogenous p16^{INK4a} promoter. Chromatin immunoprecipitation assays were performed using antiserum that recognizes both Ets1 and Ets2 and the precipitated DNA was examined for the presence of p16^{INK4a} promoter sequences by polymerase chain reaction (PCR) and Southern blotting. The data provided clear evidence for an interaction between Ets1/Ets2 and the relevant domain of the p16^{INK4a} promoter in late passage HDFs (Fig. 2a). Moreover, ectopic expression of oncogenic Ras increased Ets1/Ets2 binding to the p16^{INK4a} promoter in young HDFs (Fig. 2b). Conversely, when E2DBD sequences were expressed in primary HDFs, using a recombinant retrovirus¹⁷, they caused a significant reduction in the level of endogenous p16^{INK4a} without affecting the level of MEK, used here as an internal control (Fig. 2c). A similar result was obtained using the pIRES2-EGFP bicistronic vector to transiently express E2DBD in SVts8 cells (Fig. 2d). As

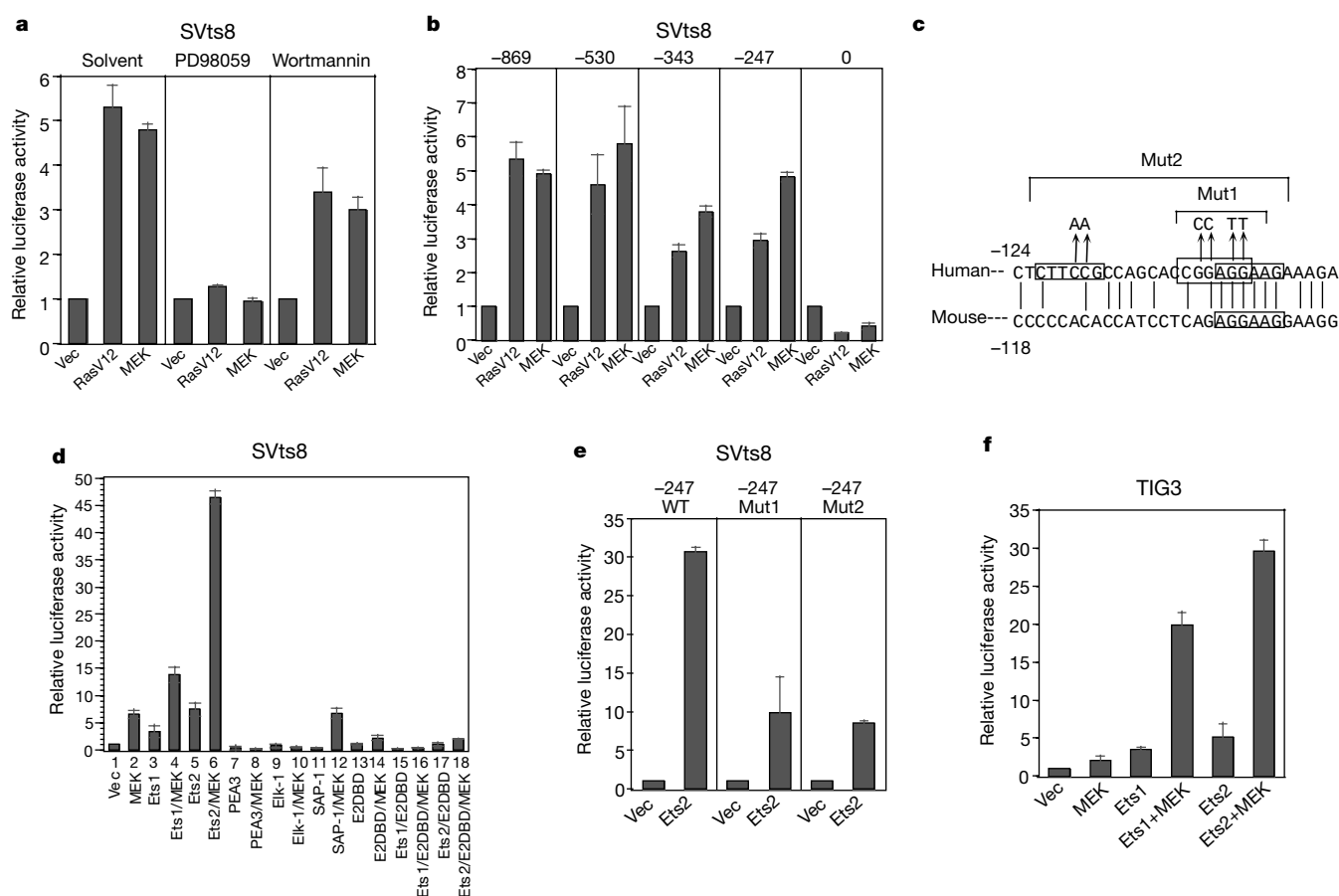


Figure 1 Modulation of p16^{INK4a} promoter activity. **a**, Activation of the p16^{INK4a} promoter (-869 construct) by H-RasV12 or activated MEK in SVts8 cells and effects of PD98059 or wortmannin. **b**, Progressively deleted p16^{INK4a} promoter constructs³ assayed with or without H-RasV12 or activated MEK. **c**, Ets-binding sequences in the human and mouse p16^{INK4a} promoters (accession numbers X94154 and U47018) and the Mut1/Mut2

mutations. **d**, Effects of Ets family proteins on p16^{INK4a} promoter activity (-869 construct) with and without MEK. **e**, Effects of Mut1 and Mut2 mutations on response to Ets2. **f**, Activation of the p16^{INK4a} promoter (-869 construct) by Ets2 and MEK in primary HDFs. WT, wild type.

further proof of the *in vivo* consequences of these effects, early passage HDFs were infected with a recombinant adenovirus encoding Flag-tagged Ets2. Control cells infected with the empty vector grew normally to confluence within 5 days, whereas the Ets2-expressing cells stopped dividing within 2–3 days and never reached confluence (Fig. 2f and g). Furthermore, Ets2-expressing cells contained increased levels of endogenous p16^{INK4a} (Fig. 2e) and stained positively for senescence-associated β -galactosidase activity¹⁹ (see Supplementary Information). Similar results were obtained with early passage HDFs infected with recombinant retroviruses encoding Ets1 or Ets2 (data not shown).

The Ets family of transcription factors can be negatively regulated by members of the Id family of helix–loop–helix (HLH) proteins²⁰, better known as dominant negative inhibitors of lineage specific bHLH proteins²¹. As loss of Id function has been linked to replicative senescence²², and mouse embryo fibroblasts (MEFs) derived from Id1-deficient mice express high levels of p16^{INK4a} (ref. 23), it seemed possible that Id1 and Ets have antagonistic influences on p16^{INK4a} expression. First we determined that labelled Ets2 binds directly to GST-Id1 *in vitro* (Fig. 3a). Similarly, after co-transfecting cells with plasmids encoding Id1 and Flag-tagged Ets2, it was possible to demonstrate co-precipitation of the two proteins with antibodies against either Id1 or Flag (Fig. 3b). We next asked whether co-transfection of Id1 would affect the ability of Ets2 and/or MEK to activate the p16^{INK4a} promoter. With the –869 reporter

construct, Id1 reduced the activation by Ets2 and/or MEK in a dose-dependent manner (Fig. 3c). Although a similar effect was noted using the –247 promoter construct, co-expression of Id1 had no appreciable impact on the activity of the Mut1 and Mut2 versions of the promoter in which the Ets-binding site had been mutated (Fig. 3d). By directly opposing Ets2 function, Id1 might serve to counterbalance the activation of the p16^{INK4a} promoter mediated by Ras–Raf–MEK.

To explore this possibility, primary HDFs were infected with recombinant retroviruses encoding Id1 and H-RasV12. As others have shown⁶, introduction of Ras resulted in a substantial increase in the amount of p16^{INK4a} expressed in the control cells but prior expression of Id1 reduced the magnitude of this response (Fig. 4a). In normal circumstances, p16^{INK4a} levels increase substantially in senescent HDFs, whereas the levels of Id1 decline^{23,22}. As well as confirming these trends (Fig. 4b), we found that the Ets2 levels also declined, and we did not detect phosphorylated MEK in the senescent HDFs (Fig. 4b), suggesting that Ras-mediated activation of Ets2 cannot be responsible for the marked accumulation of p16^{INK4a} at the end of cellular lifespan. Although it was impossible to quantitate the relative amounts of Ets1 and Ets2, because they were detected by different antisera, it was clear that the loss of Ets2 from senescent cells was in part compensated for by a reciprocal increase in Ets1 expression (Fig. 4b). We therefore propose a model in which MEK and Ets2 are the main transmitters of Ras signalling to p16^{INK4a} in young cells, whereas Ets1 assumes this role when the Ras–Raf–MEK pathway is shut down, aided by the concomitant down-regulation of Id1 (see Supplementary Information).

Taken together, our data point to critical roles for Ets and Id proteins in regulating p16^{INK4a} expression in response to different oncogenic stimuli and draw a distinction between the mechanisms operating in young and old HDFs. Unless provoked by oncogenic

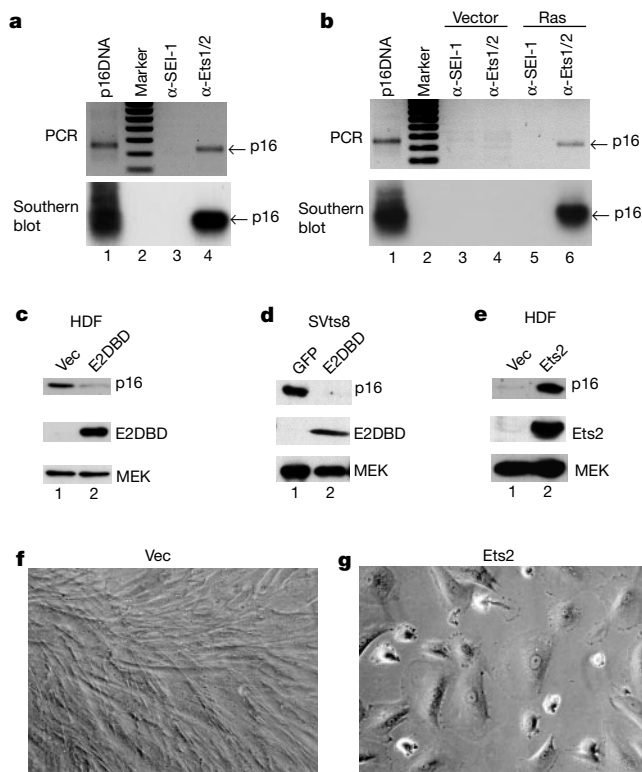


Figure 2 Effects of Ets proteins on endogenous p16^{INK4a}. **a**, Chromatin immunoprecipitates (ChIP) from senescent HDFs using Ets1/2 or SEI-1 antisera²⁸ were analysed by PCR using primers flanking the Ets sites in the p16^{INK4a} promoter (upper panel) and hybridized with a p16^{INK4a} specific probe (lower panel). **b**, Equivalent ChIP assays in young HDFs infected with control or recombinant retroviruses encoding H-RasV12. **c**, Downregulation of endogenous p16^{INK4a} in HDFs by retrovirus encoding dominant negative E2DBD. **d**, Downregulation of p16^{INK4a} in SVts8 cells transfected with E2DBD-IRES-GFP. **e**, Upregulation of p16^{INK4a} in young HDFs infected with adenovirus encoding Ets2. Levels of endogenous p16^{INK4a} and exogenous Ets2 were analysed by immunoblotting with antibodies against either p16^{INK4a} or Flag. **f** and **g**, Photomicrographs of HDFs infected with empty vector (**f**) or Ets2 adenovirus (**g**) at 5 days after infection.

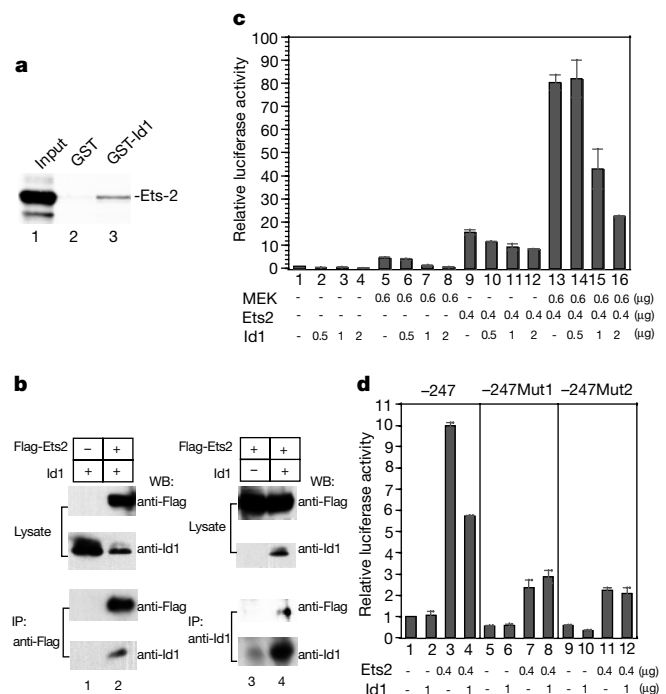


Figure 3 Id1 interacts with Ets2 and blocks its effects on the p16^{INK4a} promoter. **a**, Binding of ³⁵S-labelled Ets2 to GST (lane 2) or GST–Id1 fusion protein (lane 3). **b**, Association of Flag-tagged Ets2 and Id1 in transfected 293T cells. Lysates were analysed directly or immunoprecipitated with Flag or Id1 antibodies as indicated. **c**, Dose-dependent ability of Id1 to block activation of the p16^{INK4a} promoter by Ets2 and activated MEK in SVts8 cells. **d**, Sensitivity of the Mut1, Mut2 and –247 promoter constructs to different amounts of Ets2 and Id1 in SVts8 cells.

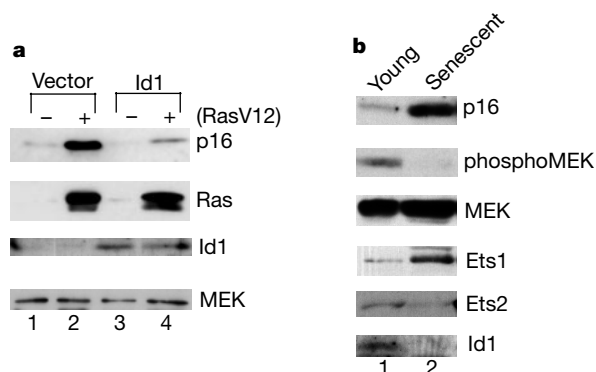


Figure 4 Expression of Id1 and Ets proteins and their effects on endogenous p16^{INK4a} levels. **a**, Early passage HDFs infected with a recombinant retrovirus encoding Id1 or empty vector were superinfected with a retrovirus encoding H-RasV12 or control vector. After seven days, p16^{INK4a}, Ras, Id1 and MEK levels were analysed by immunoblotting. **b**, Relative levels of Ets1, Ets2 and Id1 in young (40 population doublings) and senescent (72 population doublings) TIG-3 cells.

Ras, young HDFs express relatively low levels of p16^{INK4a}, presumably because signalling through the predominant Ets protein, Ets2, is counterbalanced by Id1. Introduction of oncogenic Ras would override this steady state by promoting the aberrant phosphorylation of Ets2 (model shown in Supplementary Information). In senescent cells, however, where Ras–Raf–MEK signalling is attenuated and Ets2 levels are low, upregulation of p16^{INK4a} depends on the increased expression of Ets1, in the absence of any interference by Id1. These ideas are consistent with recent reports that Id1 can collaborate with viral oncoproteins in the bypass of senescence²⁴ and that Id1 can extend the lifespan of primary keratinocytes²⁵. However, it is clear that all aspects of p16^{INK4a} regulation cannot be explained by effects described here and that the promoter is subject to multiple levels of control^{13,26}. As well as establishing the hierarchy and interplay of these regulatory mechanisms, it will be important to determine whether additional proteins are involved in the regulation of p16^{INK4a} by Ets1 and Ets2. □

Methods

Cells

TIG3 and Hs68 strains of HDFs were rendered sensitive to infection by ecotropic retroviruses as described¹⁷ and infected with recombinant retroviruses encoding H-RasV12 (in pBABEbleo) or Id1 (in pBABEpuro). Pools of drug-resistant cells were analysed 7 days after infection. In some experiments, SVt8 cells were transfected with pIRES-EGFP bicistronic vector (Clontech) encoding the Ets2 DNA-binding domain (E2DBD). Green fluorescent protein (GFP) positive cells were isolated by fluorescence-activated cell sorting.

Promoter assays

Previously characterized p16^{INK4a} promoter-reporter plasmids were assayed in SVt8 cells as described³ or in TIG3 cells transfected using FuGENE6 (Boehringer Mannheim). Effector plasmids were co-transfected as indicated in the figures, along with a standard amount of the MMLV-lacZ control plasmid. Cells were harvested 48 h after transfection and assayed for luciferase and β -galactosidase⁷. Luciferase activities were normalized to the corresponding β -galactosidase activity. For the various effector plasmids, relative promoter activities were calculated using the empty vector control as the reference.

Chromatin immunoprecipitations

ChIP assays were performed using a modification of a published protocol²⁷. After immunoprecipitation with a polyclonal antiserum against Ets1 or Ets2 (no. 57 from J. Ghysdael) or an irrelevant control protein (SEI-1)²⁸, the recovered DNA was analysed by PCR with primers flanking the putative Ets site: 5'-TGCTCGGGGTTAATAGCACC-3' and 5'-CTCCATGCTGCCCGCCG-3'. The products were analysed by gel electrophoresis and Southern blotting with a p16^{INK4a} genomic probe.

Protein analysis

Immunoblotting and immunoprecipitation were performed as described²⁸ with primary antibodies against p16^{INK4a} (DCS50, JC8 and DPAR12), MEK1/2 (NEB no. 9122),

Phospho-MEK1/2 (NEB no. 9121S), Ets1 (C-20 and N-276; Santa Cruz), Ets2 (C-20; Santa Cruz), Id1 (C-20; Santa Cruz), Ras (Calbiochem Ab-4) and Flag M2 (Sigma A-1205). The *in vitro* binding assays were performed as described²⁰.

Received 11 August; accepted 28 November 2000.

- Serrano, M., Hannon, G. J. & Beach, D. A new regulatory motif in cell cycle control causing specific inhibition of cyclin/CDK4. *Nature* **366**, 704–707 (1993).
- Alcorta, D. A. et al. Involvement of the cyclin-dependent kinase inhibitor p16 (INK4a) in replicative senescence. *Proc. Natl Acad. Sci. USA* **93**, 13742–13747 (1996).
- Hara, E. et al. Regulation of p16^{CDKN2} expression and its implications for cell immortalization and senescence. *Mol. Cell. Biol.* **16**, 859–867 (1996).
- Loughran, O. et al. Association of CDKN2A/p16^{INK4a} with human head and neck keratinocyte replicative senescence: relationship of dysfunction to immortality and neoplasia. *Oncogene* **13**, 561–568 (1996).
- Reznikoff, C. A. et al. Elevated p16 at senescence and loss of p16 at immortalization in human papillomavirus 16 E6, but not E7, transformed human uroepithelial cells. *Cancer Res.* **56**, 2886–2890 (1996).
- Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic *ras* provokes premature cell senescence associated with accumulation of p53 and p16^{INK4a}. *Cell* **88**, 593–602 (1997).
- Lin, A. W. et al. Premature senescence involving p53 and p16 is activated in response to constitutive MEK/MAPK mitogenic signaling. *Genes Dev.* **12**, 3008–3019 (1998).
- Zhu, J., Woods, D., McMahon, M. & Bishop, J. M. Senescence of human fibroblasts induced by oncogenic Raf. *Genes Dev.* **12**, 2997–3007 (1998).
- Ruas, M. & Peters, G. The p16^{INK4a}/CDKN2A tumor suppressor and its relatives. *Biochim. Biophys. Acta* **1378**, 115–117 (1998).
- Graves, B. J. & Petersen, J. M. Specificity within the ets family of transcription factors. *Adv. Cancer Res.* **75**, 1–55 (1998).
- Benezra, R., Davis, R. L., Lockshon, D., Turner, D. L. & Weintraub, H. The protein Id: a negative regulator of helix-loop-helix DNA binding proteins. *Cell* **61**, 49–59 (1990).
- Sherr, C. J. & Roberts, J. M. CDK inhibitors: positive and negative regulators of G1-phase progression. *Genes Dev.* **13**, 1501–1512 (1999).
- Hayflick, L. & Moorhead, P. S. The serial cultivation of human diploid cell strains. *Exp. Cell Res.* **25**, 585–621 (1961).
- Lundberg, A. S., Hahn, W. C., Gupta, P. & Weinberg, R. A. Genes involved in senescence and immortalization. *Curr. Opin. Cell Biol.* **12**, 705–709 (2000).
- Harley, C. B., Futcher, A. B. & Greider, C. W. Telomeres shorten during ageing of human fibroblasts. *Nature* **345**, 458–460 (1990).
- Sedivy, J. M. Can ends justify the means: telomeres and the mechanisms of replicative senescence and immortalization in mammalian cells. *Proc. Natl Acad. Sci. USA* **95**, 9078–9081 (1998).
- McConnell, B. B., Starborg, M., Brookes, S. & Peters, G. Inhibitors of cyclin-dependent kinases induce features of replicative senescence in early passage human diploid fibroblasts. *Curr. Biol.* **8**, 351–354 (1998).
- Foos, G., Garcia-Ramirez, J. J., Galang, C. K. & Hauser, C. A. Elevated expression of Ets2 or distinct portions of Ets2 can reverse ras-mediated cellular transformation. *J. Biol. Chem.* **273**, 18871–18880 (1998).
- Dimri, G. P. et al. A novel biomarker identifies senescent human cells in culture and in aging skin *in vivo*. *Proc. Natl Acad. Sci. USA* **92**, 9363–9367 (1995).
- Yates, P. R., Atherton, G. T., Deed, R. W., Norton, J. D. & Sharrocks, A. D. Id helix-loop-helix proteins inhibit nucleoprotein complex formation by the TCF ETS-domain transcription factors. *EMBO J.* **18**, 968–976 (1999).
- Massari, M. E. & Murre, C. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Mol. Cell. Biol.* **20**, 429–440 (2000).
- Hara, E. et al. Id-related genes encoding helix-loop-helix proteins are required for G1 progression and are repressed in senescent human fibroblasts. *J. Biol. Chem.* **269**, 2139–2145 (1994).
- Lyden, D. et al. Id1 and Id3 are required for neurogenesis, angiogenesis and vascularization of tumour xenografts. *Nature* **401**, 670–677 (1999).
- Hara, E. et al. The helix-loop-helix protein Id-1 and a retinoblastoma protein binding mutant of SV40 T antigen synergize to reactivate DNA synthesis in senescent human fibroblasts. *Dev. Genet.* **18**, 161–172 (1996).
- Alani, R. M. et al. Immortalization of primary human keratinocytes by the helix-loop-helix protein, Id-1. *Proc. Natl Acad. Sci. USA* **96**, 9637–9641 (1999).
- Jacobs, J. J. L., Kieboom, K., Marino, S., DePinho, R. A. & van Lohuizen, M. The oncogene and Polycomb-group gene *bmi-1* regulates cell proliferation and senescence through the *ink4a* locus. *Nature* **397**, 164–168 (1999).
- Takahashi, Y., Rayman, J. B. & Dynlacht, B. D. Analysis of promoter binding by the E2F and pRB families *in vivo*: distinct E2F proteins mediate activation and repression. *Genes Dev.* **14**, 804–816 (2000).
- Sugimoto, M. et al. Regulation of CDK4 activity by a novel CDK4-binding protein, p34^{SEI-1}. *Genes Dev.* **13**, 3027–3033 (1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank C. A. Hauser, K. E. Boulukos, N. G. Ahn, Y. de Launoit, Y. Nagamine and J. Ghysdael for providing useful materials. We are also grateful to N. Jones, R. Treisman, J. Campisi, I. Palmero and M. Serrano for helpful discussions, and to M. Hughes and J. Barry for help in FACS. We also thank T. Tanaka for his useful suggestion in ChIP assay and S. Bagley and T. D. Allen for help in using the microscope. This work was supported by the Cancer Research Campaign, Imperial Cancer Research Fund and Nihon Schering KK.

Correspondence and requests for materials should be addressed to E.H. (e-mail: Ehara@picr.man.ac.uk).

Structural determinants for regulation of phosphodiesterase by a G protein at 2.0 Å

Kevin C. Slep^{*†‡}, Michele A. Kercher^{*‡§}, Wei He^{||},
Christopher W. Cowan^{†||}, Theodore G. Wensell^{||} & Paul B. Sigler^{*§}

^{*} Department of Molecular Biophysics and Biochemistry, and [§] Howard Hughes Medical Institute, Yale University, New Haven, Connecticut 06511, USA

^{||} Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, Texas 77030, USA

[‡] These authors contributed equally to this work

[†] Present addresses: Department of Cellular and Molecular Pharmacology and Howard Hughes Medical Institute, University of California, San Francisco, California 94143, USA (K.C.S.); Division of Neuroscience, Children's Hospital, and Department of Neurobiology, Harvard Medical School, Boston, Massachusetts 02115, USA (C.W.C.)

A multitude of heptahelical receptors use heterotrimeric G proteins to transduce signals to specific effector target molecules. The G protein transducin, G_t , couples photon-activated rhodopsin with the effector cyclic GMP phosphodiesterase (PDE) in the vertebrate phototransduction cascade. The interactions of the G_t α -subunit (α_t) with the inhibitory PDE γ -subunit (PDE γ) are central to effector activation, and also enhance visual recovery in cooperation with the GTPase-activating protein regulator of G-protein signalling (RGS)-9 (refs 1–3). Here we describe the crystal structure at 2.0 Å of rod transducin α -GDP- AlF_4^- in complex with the effector molecule PDE γ and the GTPase-activating protein RGS9. In addition, we present the independently solved crystal structures of the RGS9 RGS domain both alone and in complex with $\alpha_{t/i1}$ -GDP- AlF_4^- . These structures reveal insights into effector activation, synergistic GTPase acceleration, RGS9 specificity and RGS activity. Effector binding to a nucleotide-dependent site on α_t sequesters PDE γ residues implicated in PDE inhibition, and potentiates recruitment of RGS9 for hydrolytic transition state stabilization and concomitant signal termination.

Crystals were obtained using the RGS9 RGS domain (residues 276–422), an NΔ25 $\alpha_{t/i1}$ chimaera, and PDE γ (residues 46–87). The NΔ25 $\alpha_{t/i1}$ chimaera is identical to α_t except for residues 216–294 which are replaced with the corresponding homologous region of

α_{i1} (residues 220–298)⁴. The $\alpha_{t/i1}$ chimaera interacts with PDE γ_{46-87} and displays an augmented rate of RGS9₂₇₆₋₄₂₂/PDE γ_{46-87} -mediated GTP hydrolysis⁵ (data not shown). The constructs that we used reflect functional recombinant components that yield diffraction-quality crystals. All three structures were solved by multiwavelength anomalous dispersion (MAD) phasing using selenomethionine (Se-Met)-substituted RGS9. Diffraction data and final refinement statistics are presented in Table 1.

The structure of the $\alpha_{t/i1}$ -GDP- AlF_4^- -PDE γ -RGS9 complex is depicted in Fig. 1. Secondary structure observed in the heterotrimeric complex is diagrammed in Fig. 2. The $\alpha_{t/i1}$ subunit is composed of two domains: a Ras-like GTPase and an α -helical domain. Previous structural studies^{6,7} identified three switch regions (I–III) in α_t that undergo nucleotide-dependent conformational changes (Fig. 2). The overall structure of the $\alpha_{t/i1}$ subunit and its switch region architecture most closely resembles the transition state analogue structure α_t -GDP- AlF_4^- (ref. 8). The RGS9 RGS domain is structurally homologous to the RGS domains of RGS4 (ref. 9), axin¹⁰ and G alpha interacting protein (GAIP)¹¹. Nine helices are configured into two subdomains: an amino- and carboxy-terminal region (helices α_1 , α_2 , α_3 , α_8 and α_9), and a prototypical right-handed, antiparallel four-helix bundle (helices α_4 , α_5 , α_6 and α_7). The two subdomains are united by a hydrophobic interface. All independently refined structures of the RGS9 RGS domain show similar secondary structure with minor variations in helix length and in the occurrence of single turn helices in loops not used in G_α binding. PDE γ_{46-87} comprises three short α -helices and an N-terminal loop region that originates near the C terminus and winds over helices α_1 and α_2 . Helices α_1 and α_2 adopt a relative orientation that presents the critical residue of PDE γ , Trp 70, to $\alpha_{t/i1}$.

PDE γ engages the $\alpha_{t/i1}$ switch II/ α_3 cleft in a hydrophobic interlock that buries 1,500 Å² of solvent-accessible surface (Fig. 3a–f). Interactions are shown in Fig. 3b. Whereas binding involves primarily hydrophobic interactions with an electrostatically neutral region of $\alpha_{t/i1}$, the PDE γ interface has a relatively negative charge distribution (Fig. 4). Trp 70 of PDE γ has a key role in the $\alpha_{t/i1}$ interaction. Transgenic and *in vitro* studies using the PDE γ mutant Trp70Ala confirm the role of Trp 70 in $\alpha_{t/i1}$ binding and GTPase-activating protein (GAP) potentiation^{12,13} (Fig. 3b). The Trp70Ala mutation slows the flash response recovery rate, decreases affinity for α_t fivefold, and fails to augment GAP activity. In the $\alpha_{t/i1}$ -GDP- AlF_4^- -PDE γ -RGS9 structure, Trp 70 binds the switch II/ α_3 cleft, forming hydrophobic contacts with seven residues from

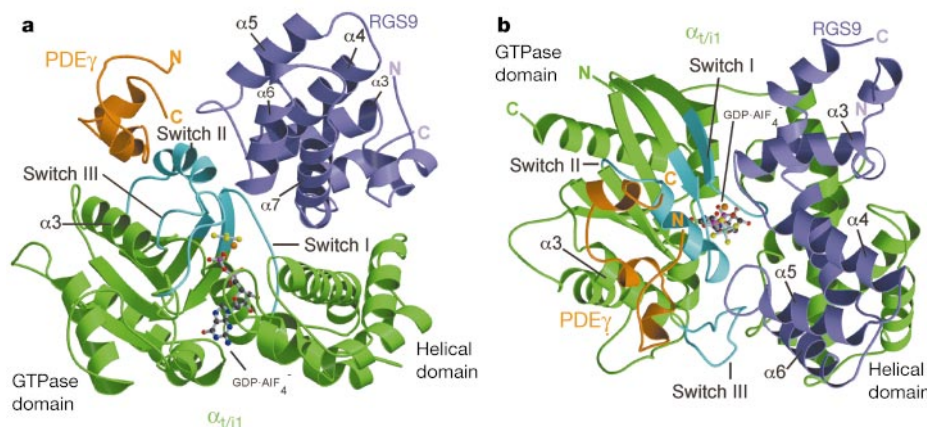


Figure 1 Structure of the $\alpha_{t/i1}$ -GDP- AlF_4^- -PDE γ -RGS9 heterotrimeric complex with $\alpha_{t/i1}$ shown in green, the switch regions of $\alpha_{t/i1}$ shown in blue, PDE γ shown in orange and RGS9 shown in lavender. **a**, Ribbon drawing of the heterotrimeric complex with GDP and AlF_4^- shown as ball-and-stick models. Carbon, nitrogen, oxygen, phosphorus, fluoride and

aluminium atoms are coloured black, blue, red, magenta, yellow-green and green, respectively. Mg^{2+} is represented by an orange sphere. **b**, Ribbon drawing of the complex as in **a**, rotated 90° about the horizontal axis.

switch II and $\alpha 3$ (Fig. 3b, 3d). A hydrogen bond between Ser 248 of $\alpha_{t/i1}$ and the indole nitrogen restricts Trp 70 and drives the side chain into contact with $\alpha_{t/i1}$ residues Arg 204 and Trp 207 (Fig. 3e). In exchange, Trp 254 of $\alpha_{t/i1}$ extends towards the PDE γ $\alpha 3$ helix, forming hydrophobic and polar interactions with five PDE γ residues. In contrast to the RGS9/ $\alpha_{t/i1}$ interface discussed below, the PDE γ / $\alpha_{t/i1}$ interface is nearly devoid of solvent (Fig. 3b).

The heterotrimeric structure confirms the switch II/ $\alpha 3$ cleft as a primary determinant in PDE γ binding^{4,14}. Alanine scanning mutagenesis of α_t , coupled with binding and GTPase assays, implicated Arg 201, Arg 204, Trp 207, Ile 208 of switch II and His 244 of $\alpha 3$ as residues that bind PDE γ ¹⁵. The structure is also in accord with an α_t mutation that disrupts the switch II/ $\alpha 3$ cleft, prevents PDE γ binding and underlies the molecular basis for Nougaret night blindness¹⁶. The structure is not consistent with crosslinking results^{17,18} implicating binding regions outside the switch II/ $\alpha 3$ cleft.

The $\alpha_{t/i1}$ /PDE γ interaction reveals an emerging paradigm in heterotrimeric G-protein/effector coupling. Temporally regulated effector activation requires a GTP-dependent $G\alpha$ -binding site, restricting effector binding to $G\alpha$ sites that are switch region inclusive. In addition, the binding site must confer effector specificity. Switch II and $\alpha 3$, together, meet both requirements without compromising the RGS-binding site (Fig. 3f). PDE γ binding involves a number of key residues in the $\alpha_{t/i1}$ switch II/ $\alpha 3$ cleft: Arg 204, Trp 207, Phe 211, Leu 245 and Trp 254 (Fig. 5a).

The structure of adenylyl cyclase VC₁-IIC₂ in complex with α_s -GTP γ S¹⁹ reveals a strong similarity in effector binding (Fig. 5b). In PDE γ binding, Trp 254 of α_t makes hydrophobic contacts with

Leu 78 and Leu 81 of PDE γ , and forms a hydrogen bond between its indole nitrogen and the Leu 76 backbone. In adenylyl cyclase binding, the corresponding α_s residue, Trp 281, makes concerted hydrophobic contacts with Phe 379 of VC₁ and His 989 of IIC₂. A hydrogen bond from Glu 917 to the indole nitrogen of Trp 281 directly parallels the interaction of Leu 76 of PDE γ with Trp 254. Analogous to the key PDE γ residue Trp 70, the adenylyl cyclase IIC₂ residue Phe 991 is inserted into the switch II/ $\alpha 3$ cleft forming homologous hydrophobic contacts with residues Arg 231, Trp 234, Phe 238 and Leu 272 of α_s . Similar to the effects of the PDE γ mutation Trp70Ala, the IIC₂ mutation, Phe991Ala, results in decreased α_s -mediated stimulation of adenylyl cyclase activity²⁰. It is likely that the $G\alpha$ switch II/ $\alpha 3$ cleft is involved in interactions with other effector molecules, particularly where effector activation is found coupled with RGS activity.

By sequestering PDE γ residues implicated in PDE $\alpha\beta$ inhibition, the $\alpha_{t/i1}$ -GDP- AlF_4 -PDE γ -RGS9 structure reveals the mechanism of PDE activation. Proteolysis experiments²¹ and deletion mutagenesis²² have localized the inhibitory region of PDE γ to the C-terminal five amino acids. Ile 87, the C-terminal residue of PDE γ , forms contacts with switch II residues Ile 208, His 209 and Lys 212. As a result, the PDE γ C-terminal region is pulled towards switch II (Fig. 1) and is partially shielded by Phe 50 of PDE γ (Fig. 3f). These interactions constrain the PDE γ C terminus and prevent access to PDE $\alpha\beta$. In this state, PDE $\alpha\beta$ proceeds with the catalysis of cGMP, propagating and amplifying the visual signal until GTP hydrolysis inactivates α_t , and the PDE γ C terminus returns to its inhibitory site on PDE $\alpha\beta$.

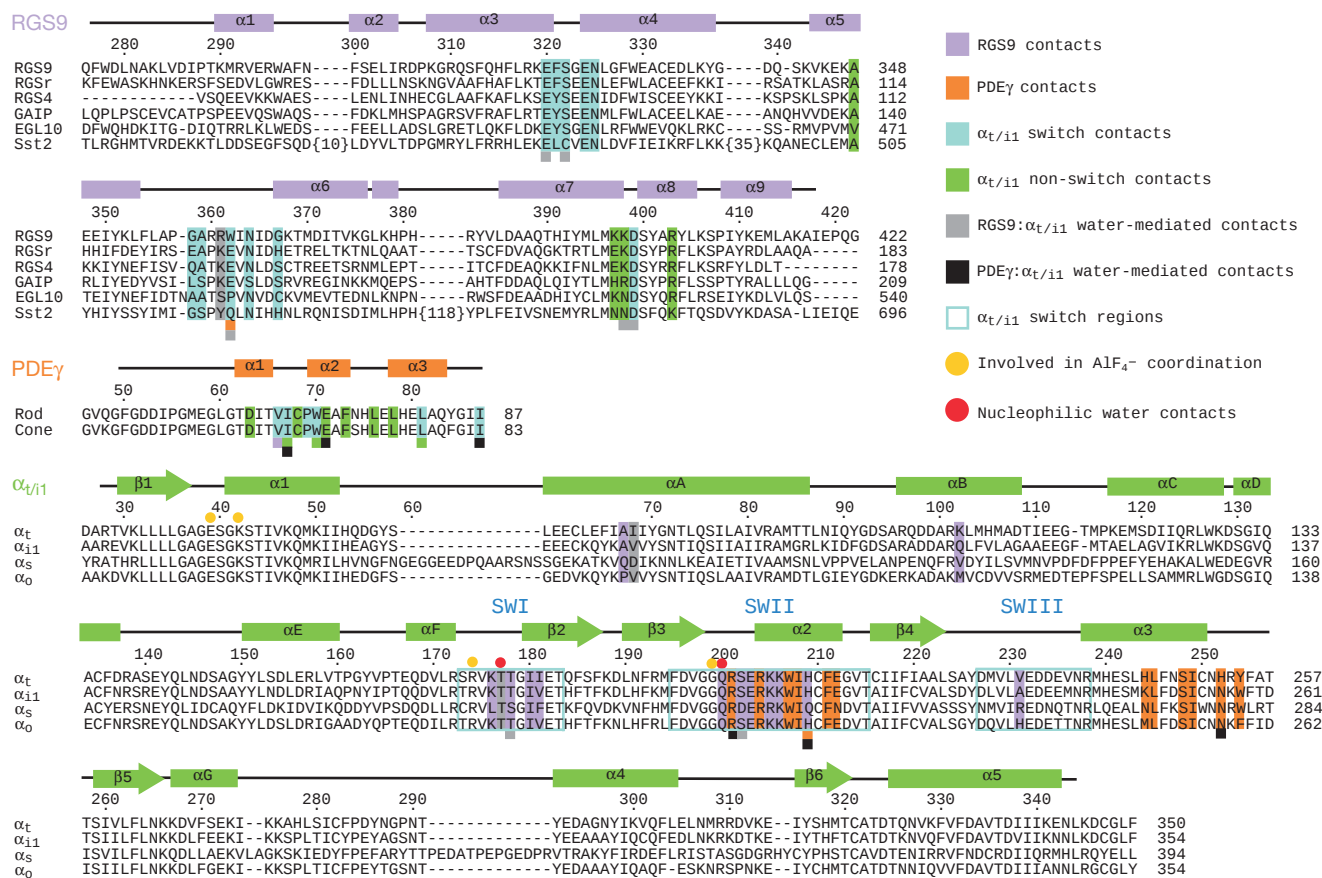


Figure 2 Sequence alignment and secondary-structure features of RGS domains, PDE γ subunits and $G\alpha$ subunits. Secondary-structure elements from the $\alpha_{t/i1}$ -GDP- AlF_4 -PDE γ -RGS9 complex are indicated above the alignments by rectangles (α -helices) and arrows (β -strands). Residue numbers are shown for RGS9, rod PDE γ and $\alpha_{t/i1}$. Residues involved in more than one of the above interactions have coloured blocks

below the alignment indicative of the additional interaction(s). Specific interactions are less than 4 Å apart. GenBank accession numbers for proteins from top to bottom are 046469, P97428, P49799, NP_005864, P49809, P11972, P04972, P22571, P04695, P04898, P04896, P08239.

Nucleotide hydrolysis produces a marked conformational change in switch II that distorts the α_i effector site. The crystal structures of α_i -GDP⁷ and $\alpha_{i/i1}$ -GDP- $\beta\gamma_i$ (ref. 23) show that, after GTP hydrolysis, switch II undergoes a torsional movement away from the α_3 helix. In the $\alpha_{i/i1}$ -GDP- $\beta\gamma_i$ structure, α_2 is rotated $\sim 90^\circ$ about its helical axis (Fig. 5c). The $\alpha_{i/i1}$ residues Ser 248 and Arg 204, which are used in PDE γ binding (Fig. 5a), now form an intimate interaction with one another (Fig. 5c). Trp 207 and Phe 211, previously exposed in the PDE γ -binding site (Fig. 5a), are now positioned on the opposite side of switch II and are used in a key set of hydrophobic contacts with β_i residues Trp 99, Leu 117 and Tyr 145 (Fig. 5c). In effect, nucleotide hydrolysis and G $\alpha\beta\gamma$ heterotrimer formation abrogates the G α -binding site used for PDE γ and adenylyl cyclase interactions.

Although our construct, PDE γ_{46-87} , contains the regions involved in nucleotide-dependent binding to α_i , GAP potentiation and PDE $\alpha\beta$ inhibition, a central polycationic region not present in our construct has been implicated in nucleotide-independent binding to α_i . The proposed binding site for the polycationic region is the α_i α_3 - β_5 region⁴. A likely trajectory from the first residue modelled (PDE γ Phe 50) to the α_3 - β_5 zone would bring the PDE γ N-terminal region through a groove created by the RGS9 α_3/α_4 loop and the α_3 /C-terminal region of PDE γ . This route would further restrain the PDE γ C terminus and augment the PDE γ /RGS9 interface.

PDE γ inactivates transducin by potentiating RGS9 GAP activity^{2,5}. The heterotrimeric structure reveals four possible means that PDE γ uses to enhance RGS9 activity. First, PDE γ expands the switch II/ α_3 cleft and positions switch II in the

transition state conformation most favoured for RGS9 binding. Switch II residues involved in PDE γ interactions dovetail with switch II residues involved in RGS9 interactions (Fig. 2). As a result, switch II is utilized to bind PDE γ and RGS9 cooperatively.

Second, PDE γ perturbs key transducin residues involved in RGS9 binding and GTP hydrolysis. PDE γ constrains $\alpha_{i/i1}$ Arg 204 through contacts with PDE γ Ile 67, Pro 69 and Trp 70 (Fig. 3e). As observed in the α_i -GTP γ S structure, Arg 204 forms two key hydrogen bonds: one from the N ϵ nitrogen to the Gly 199 carbonyl and a second α -helical hydrogen bond from the Arg 204 peptide amide nitrogen to the Gln 200 backbone carbonyl. Thus, Arg 204 effectively positions the backbone of transducin Gln 200, a key residue for GTP hydrolysis and the RGS9 GAP mechanism (discussed below). The PDE γ -mediated perturbation of the Gln 200 backbone orients the side chain for optimal interaction with the RGS9 residue Asn 364. In addition, repositioning the Gln 200 backbone may facilitate the release of the nucleophilic water which, in the α_i -GTP γ S structure, is bound to the Gln 200 backbone amide nitrogen⁶.

The third PDE γ interaction mediates stabilization of the hydrolytic intermediate (Fig. 3e). Trp 70 of PDE γ conformationally restricts Trp 207 of $\alpha_{i/i1}$, stabilizing the hydrogen bond that Trp 207 forms between its indole nitrogen and the Gly 198 carbonyl. This constraint on the Gly 198/Gly 199 peptide bond orients the backbone amide nitrogen in a conformation that stabilizes the hydrolytic intermediate. Thus, PDE γ and RGS9 cooperate in stabilizing the transition state.

Last, a small RGS9/PDE γ interface increases the available RGS9-binding site. A hydrophobic interaction between Trp 362 of RGS9 and Val 66 of PDE γ (Fig. 3a), coupled with local electrostatic charge

Table 1 Crystallographic data, phasing and refinement

RGS9						
Space group: C2	Cell dimensions: $a = 67.3 \text{ \AA}$, $b = 71.8 \text{ \AA}$, $c = 34.6 \text{ \AA}$, $\alpha = 90^\circ$, $\beta = 96.1^\circ$, $\gamma = 90^\circ$				Phasing power†	
Data Statistics:	$d_{\min} (\text{\AA})$	Completeness (%)	I/σ	$R_{\text{sym}}(\%)^*$	+ Friedel mate	- Friedel mate
Se-Met λ_1 (1.0283 \AA)	2.13	97.1 (91.8)	29.2 (10.3)	4.0 (11.3)	0.87 (0.62)	1.10 (0.77)
Se-Met λ_2 (0.9795 \AA)	2.04	98.0 (97.3)	27.1 (8.4)	4.2 (13.7)	1.11 (0.62)	3.14 (1.71)
Se-Met λ_3 (0.9793 \AA)	2.04	98.0 (98.2)	16.4 (16.1)	5.2 (11.1)	1.14 (0.97)	2.22 (1.61)
Se-Met λ_4 (0.9351 \AA)	1.94	97.7 (91.9)	23.5 (3.1)	5.1 (30.1)	Reference	1.93 (0.87)
Refinement	(50 – 1.94 \AA)					
R value‡	22.1 (25.6)					
R _{free} §	24.8 (27.0)					
Mean B factor (min/max)	28.7 \AA ² (13.1 \AA ² /58.5 \AA ²)					
$\alpha_{i/i1}$-GDP-AIF₄-RGS9						
Space group: P2 ₁ 2 ₁ 2 ₁	Cell dimensions: $a = 96.8 \text{ \AA}$, $b = 115.1 \text{ \AA}$, $c = 136.5 \text{ \AA}$, $\alpha = 90^\circ$, $\beta = 90^\circ$, $\gamma = 90^\circ$				Phasing power†	
Data statistics:	$d_{\min} (\text{\AA})$	Completeness (%)	I/σ	$R_{\text{sym}}(\%)^*$	+ Friedel mate	- Friedel mate
Native (1.0281 \AA)	2.30	96.9 (79.6)	28.5 (1.6)	6.0 (52.7)		
Se-Met λ_1 (1.0281 \AA)	2.60	88.9 (43.9)	16.6 (1.5)	6.8 (49.5)	0.85 (0.54)	0.94 (0.56)
Se-Met λ_2 (0.9794 \AA)	2.60	92.1 (54.9)	16.3 (1.5)	7.4 (47.9)	1.41 (0.49)	2.09 (0.63)
Se-Met λ_3 (0.9792 \AA)	2.65	93.4 (62.7)	15.8 (1.6)	8.0 (48.5)	0.91 (0.59)	1.74 (0.65)
Se-Met λ_4 (0.9349 \AA)	>2.80	97.9 (84.8)	16.3 (1.9)	8.2 (51.6)	Reference	1.00 (0.38)
Refinement	(50 – 2.30 \AA)					
R value‡	23.1 (38.6)					
R _{free} §	26.8 (40.3)					
Mean B factor (min/max)	65.7 \AA ² (32.5 \AA ² /127.5 \AA ²)					
$\alpha_{i/i1}$-GDP-AIF₄-PDEγ-RGS9						
Space group: P2 ₁ 2 ₁ 2 ₁	Cell dimensions: $a = 91.6 \text{ \AA}$, $b = 115.9 \text{ \AA}$, $c = 134.2 \text{ \AA}$, $\alpha = 90^\circ$, $\beta = 90^\circ$, $\gamma = 90^\circ$				Phasing power†	
Data statistics:	$d_{\min} (\text{\AA})$	Completeness (%)	I/σ	$R_{\text{sym}}(\%)^*$	+ Friedel mate	- Friedel mate
Native (0.9793 \AA)	2.02	99.8 (99.7)	19.5 (4.4)	9.0 (55.3)		
Se-Met λ_1 (1.0280 \AA)	2.40	99.2 (94.6)	13.7 (2.1)	9.0 (49.3)	0.80 (0.62)	0.91 (0.69)
Se-Met λ_2 (0.9794 \AA)	2.40	99.5 (96.9)	13.3 (1.8)	9.7 (51.9)	1.28 (0.74)	1.69 (0.84)
Se-Met λ_3 (0.9792 \AA)	2.40	99.3 (95.8)	13.6 (1.9)	9.7 (50.5)	0.91 (0.69)	1.56 (0.84)
Se-Met λ_4 (0.9349 \AA)	2.60	99.7 (98.3)	11.6 (2.4)	11.5 (53.1)	Reference	0.86 (0.56)
Refinement	(50 – 2.02 \AA)					
R value‡	23.3 (31.9)					
R _{free} §	26.5 (35.1)					
Mean B factor (min/max)	33.8 \AA ² (6.2 \AA ² /103.5 \AA ²)					

Values in parentheses are for the highest resolution shells. All data sets had an average redundancy of 3.1 or higher.

* $R_{\text{sym}} = \sum_i \sum_j |I(h) - \langle I(h) \rangle| / \sum_i \sum_j I(h)$ where $I(h)$ is the integrated intensity of the i th reflection with the Miller index h and $\langle I(h) \rangle$ is the average over Friedel and symmetry equivalents.

† MAD phasing power is defined as $(|F_o - F_N|^2 / P(\phi)) / (F_N |e^{\phi} + \Delta F_N| - F_o |e^{\phi}|)^{1/2}$ where $P(\phi)$ is the experimental phase probability distribution, F_N are structure factors at the reference wavelength λ_r and ΔF_N are structure factors of + or - Friedel mates at the other designated wavelength. ΔF_N is the difference in heavy atom structure factors between two wavelengths.

‡ R value = $\sum (|F_o| - |F_c|) / \sum |F_o|$.

§ R_{free} is calculated using a 10% subset of the data which was removed randomly from the original data and excluded from refinement.

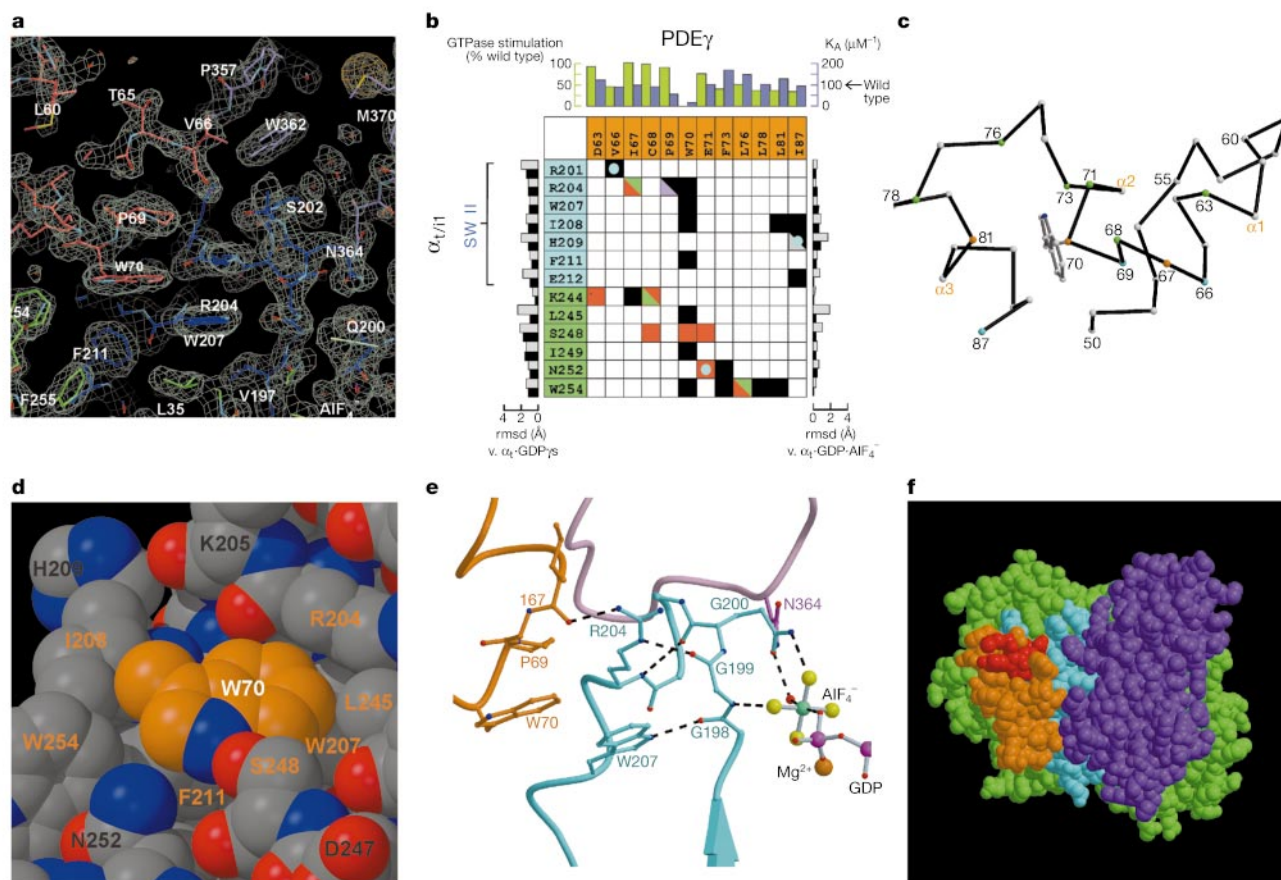


Figure 3 PDEγ/ $\alpha_{1/11}$ interactions. **a**, σ_A -weighted $2F_o - F_c$ electron density map for $\alpha_{1/11}$ -GDP·AlF₄⁻·PDEγ·RGS9 (1.5σ). Anomalous difference Fourier density (15σ) in coral. **b**, Intermolecular contact (sub 4 Å) matrix for PDEγ residues (orange) that contact $\alpha_{1/11}$ non-switch residues (green) or $\alpha_{1/11}$ switch residues (blue). Electrostatic interactions to side chains are shown in red, to main chain in green. van der Waals contacts to side chains, to main chain or to both are shown in black, grey and lavender, respectively. Water-mediated interactions are indicated by blue circles. The root-mean-square deviation (rmsd) for $\alpha_{1/11}$ contact residues (Cα atoms are in black, overall side-chain

deviation in grey) is shown for $\alpha_{1/11}$ -RGS9 versus $\alpha_{1/11}$ -GTPγS (left)⁶ and $\alpha_{1/11}$ -GDP·AlF₄⁻ (right)⁸. Results of PDEγ alanine scanning mutagenesis on GTPase stimulation (light-green bars) and K_A (dark-blue bars; ref. 13) are shown above. **c**, Cα trace of PDEγ₅₀₋₈₇. Cα atoms of residues contacting $\alpha_{1/11}$ switch II, the α3/α3-β5 loop region or both are in blue, green and orange, respectively. **d**, CPK representation of PDEγ W70 in the switch II/α3 groove with residues that contact W70 in orange. **e**, Diagram of PDEγ/switch II/RGS9 interactions coloured as in **a**. **f**, CPK representation of the heterotrimeric complex with the five C-terminal amino acids of PDEγ in red.

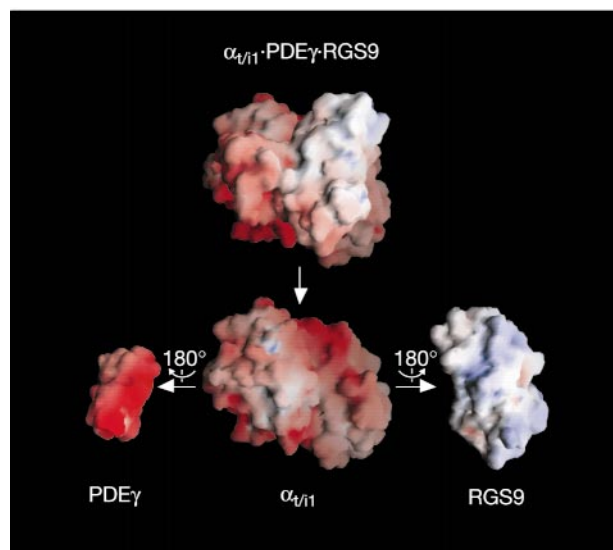


Figure 4 Solvent-accessible surface coloured according to electrostatic potential in the range $-10k_B T$ (red) to $+10k_B T$ (blue), where k_B is Boltzmann's constant, and T is the absolute temperature (in K). The trimeric complex is shown above in the same orientation

as in Fig. 3f. The orientation is maintained below except that PDEγ and RGS9 are translated horizontally and rotated 180° about the vertical axis.

complementation, supplements the RGS9-binding site by burying an additional $\sim 200 \text{ \AA}^2$ of solvent-accessible surface. The proximity of the PDE γ /RGS9 interface to the critical RGS9 GAP residue Asn 364 is likely to support GAP activity.

The RGS9/ $\alpha_{t/i1}$ interface is highly solvated and is dominated by interactions from the RGS9 $\alpha 3/\alpha 4$, $\alpha 5/\alpha 6$ and $\alpha 7/\alpha 8$ loops to the $\alpha_{t/i1}$ switch regions I and II (Fig. 6a) which buries $1,800 \text{ \AA}^2$ of solvent-accessible surface. The catalytic face of RGS9 is predominantly positively charged and complements the negatively charged $\alpha_{t/i1}$ binding surface (Fig. 4). A comparison of the interactions within the RGS4- α_{i1} complex⁹ with those within the RGS9- $\alpha_{t/i1}$ complex reveals minor deviations. The RGS9- $\alpha_{t/i1}$ complex contains fewer switch III interactions, and fewer total interactions. Analysis of all RGS9- $\alpha_{t/i1}$ protomers solved reveals a plasticity in the binding interface possibly indicative of the RGS protein's ability to recognize a range of switch conformations during GTP hydrolysis.

Comparison of the RGS9 $\alpha_{t/i1}$ structure with its independently solved component parts reveals structural changes that occur on complex formation. On binding, RGS9 causes switches I and II to move closer to each other, and induces an interswitch stabilization

that is absent in the α_t -GDP- AlF_4^- structure. RGS9 residue Asn 325 causes the side chain of Thr 178 in switch I to rotate $\sim 120^\circ$, forming interactions between Thr 178 of switch I and residues Lys 206 and Glu 203 of switch II (Fig. 6e). Thus, RGS9 binds $\alpha_{t/i1}$, serving as an external scaffold for the transition state architecture while simultaneously inducing internal switch region stabilization.

RGS9 undergoes moderate conformational changes on binding $\alpha_{t/i1}$. The $\alpha 5/6$ loop exhibits the greatest change, repositioning the critical GAP residue Asn 364 (Fig. 6b). In the RGS9 structure, Asn 364 forms hydrophobic interactions with Trp 362 (Fig. 6c). On binding $\alpha_{t/i1}$, the Asn 364/Trp 362 interaction is disrupted by the insertion of Ser 202 of $\alpha_{t/i1}$ (Fig. 6d). Serine 202 is conserved among α_i family members, but the corresponding residue in α_s is an aspartic acid. The $\alpha_{t/i1}$ mutation Ser202Asp prevents the RGS interaction and explains the immunity of α_s to RGS GAP activity²⁴. A second critical residue for the RGS GAP mechanism is Asp 399, which is partially disordered in the RGS9 structure. On binding $\alpha_{t/i1}$, Asp 399 interacts with RGS9 residues Arg 403 and Glu 320 and is positioned for switch I stabilization (Fig. 6e).

A number of RGS proteins exhibit GAP activity on α_t *in vitro*, but the specificity of RGS9 lies in its PDE γ -mediated potentiation. The effect of PDE γ on other RGS proteins is GAP suppression²⁵. The RGS9 region involved in PDE γ specificity has been localized to the $\alpha 5/6$ loop^{26,27}. This loop region interacts with both switch II and PDE γ through a unique complementation of electrostatic (RGS9 Arg 360/PDE γ Asp 52) and hydrophobic (RGS9 Trp 362/PDE γ Val 66) interactions. The RGS9 residues used in this region are unique to RGS9. In most RGS proteins, Arg 360 and Trp 362 are replaced by proline and glutamate, respectively. The repulsive effects of altering the electrostatic and hydrophobic nature of the $\alpha 5/6$ loop would inevitably force the proximal Asn 364 out of position. As discussed below, Asn 364 is a key component of the GAP mechanism; disrupting its position would have deleterious effects on GAP activity.

RGS9 increases the GTP hydrolysis rate by stabilization of the $\alpha_{t/i1}$ switch regions in their transition state conformation and orientation of the critical $\alpha_{t/i1}$ carbonyls used to position the nucleophilic water as previously suggested⁹. At switch I, interactions of RGS9 residues Asn 325 and Asp 399 with $\alpha_{t/i1}$ Thr 178 effectively position the adjacent $\alpha_{t/i1}$ Thr 177 backbone carbonyl for stabilization of the nucleophilic water (Fig. 6e). The imposed constraints position Thr 178 and the switch I backbone in a different orientation from that observed in the α_t -GDP- AlF_4^- structure (Fig. 6a).

At switch II, the orientation of RGS9 Asn 364 is dictated by hydrogen bonds to $\alpha_{t/i1}$ residues Lys 176 and Glu 203 (Fig. 6e). As positioned, the Asn 364 side-chain amide nitrogen resolves the torsional ambiguity of $\alpha_{t/i1}$ Gln 200. As a result, the Gln 200 side-chain carbonyl is positioned to stabilize the nucleophilic water while the Gln 200 side-chain nitrogen is positioned to stabilize the γ -phosphate planar intermediate.

The heterotrimeric structure reveals details of the catalytic site at 2.0 $\text{\AA}$ resolution, allowing the unambiguous conclusion that RGS proteins do not directly contribute to the active site by donating residues or through water-mediated interactions.

The $\alpha_{t/i1}$ -GDP- AlF_4^- -PDE γ -RGS9 structure reveals several important features of the mammalian rod visual cascade that facilitate signal transmission, amplification and sub-second recovery. Effector/GAP coupling is likely to be found in other high-fidelity signalling systems where a G protein must remain active until it reaches its effector target, at which point it is inactivated, ensuring signal efficiency. $\square$

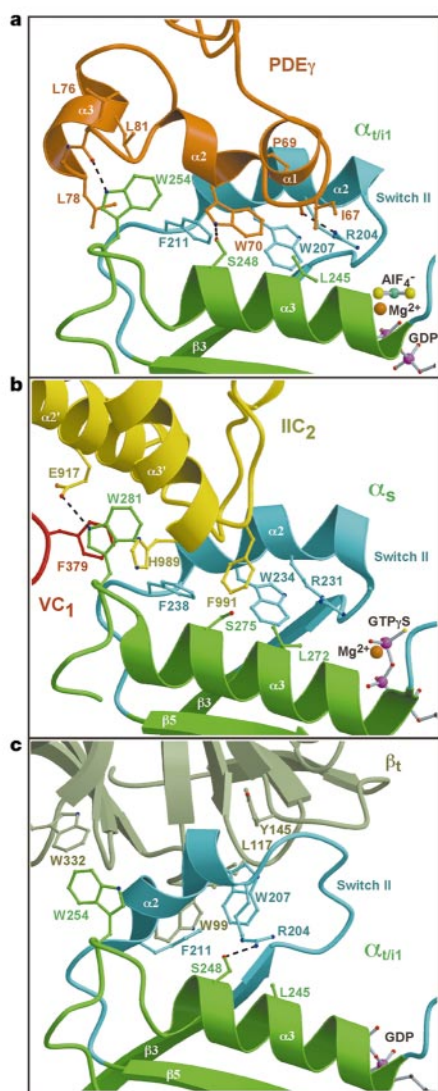


Figure 5 Interaction modes of the $G\alpha$ switch II region with $G\alpha$ effector molecules and the $G\beta\gamma$ subunit. **a**, Interaction of PDE γ with $\alpha_{t/i1}$ -GDP- AlF_4^- . The colouring scheme is the same as in Fig. 1. **b**, Interaction of adenyl cyclase constructs VC₁ (red) and IIC₂ (yellow) with α_s -GTP γ S¹⁹. The α_s colouring scheme is homologous to that used for $\alpha_{t/i1}$ in **a**. **c**, Interaction of β_t (grey-green) with $\alpha_{t/i1}$ -GDP (coloured as in **a**)²³.

Methods

Protein expression, purification and crystallization

We subcloned bovine complementary DNA encoding the RGS9 RGS domain² (residues 276–422) into pGEX-2T (Pharmacia). Native and Se-Met-substituted RGS9 proteins were expressed in *Escherichia coli* and purified using glutathione Sepharose 4B (Pharmacia).

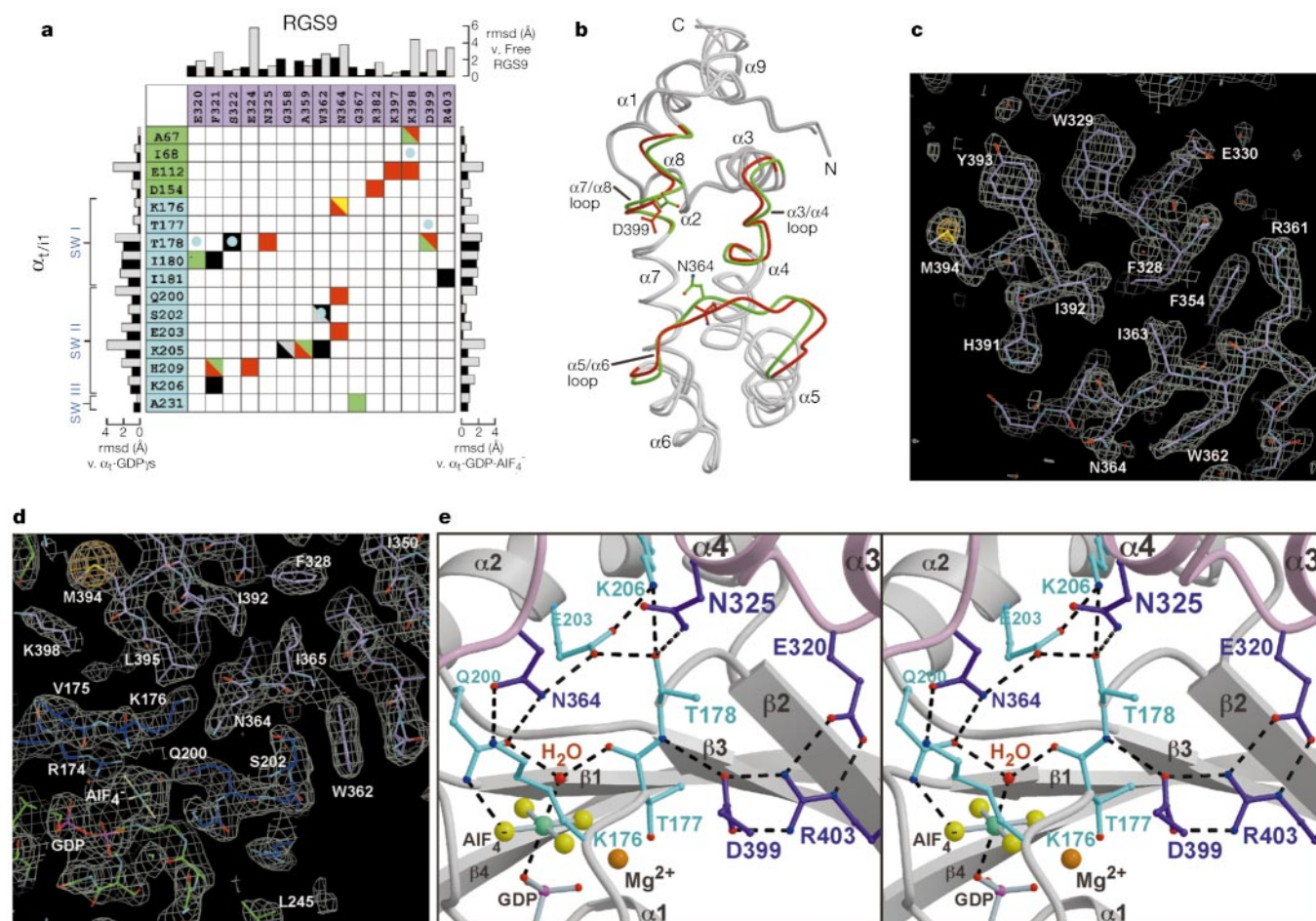


Figure 6 RGS9/ α_{i1} interactions. **a**, Intermolecular contact matrix for RGS9 residues (lavender) that contact α_{i1} residues represented as in Fig. 3b except that yellow indicates electrostatic interactions to both side-chain and main-chain atoms. **b**, C α trace comparison of RGS9 interaction loops: RGS9 alone (red) compared with RGS9 in complex

with α_{i1} (green). **c**, **d**, σ A-weighted $2F_o - F_c$ electron density map (1.5σ) and anomalous difference Fourier density (15σ , in coral) of RGS9 (**c**) and α_{i1} -GDP- AlF_4^- -RGS9 (**d**). **e**, Stereo-pair ball-and-stick model of the active site. RGS9 and α_{i1} are pink and grey, respectively.

Fusion protein was cleaved with thrombin and subjected to cation exchange chromatography. α_{i1} was prepared as described⁴, except that α_{i1} -GDP- AlF_4^- was treated with endoproteinase LysC to produce residues 26–350 of α_{i1} . We subcloned DNA encoding the bovine PDE γ subunit (GenBank accession number M28853) (residues 46–87) into pGEX-2T, expressed it in *E. coli*, and purified it using glutathione Sepharose 4B. Fusion protein was cleaved with thrombin and subjected to Q Sepharose and Superdex 75 16/60 (Pharmacia) chromatography.

The α_{i1} -GDP- AlF_4^- -RGS9 complex was formed and purified using a described protocol⁹. All crystals were obtained by hanging-drop vapour diffusion at 4 °C against 1 ml of respective mother liquor. A Se-Met RGS9 crystal was obtained by mixing 1.5 μl of RGS9 at 15 mg ml⁻¹ with 1.5 μl mother liquor (15.5% PEG 8000, 50 mM Tris pH 8.0, 5% ethylene glycol, 75 mM NaCl). The crystal was transferred to 12% PEG 8000, 50 mM Tris pH 8.0, 25% ethylene glycol, 75 mM NaCl and flash frozen. The crystal belongs to the space group C2 with one molecule per asymmetric unit. α_{i1} -GDP- AlF_4^- -RGS9 crystals (native and RGS9_{Se-Met} substituted) were obtained by mixing 1.5 μl of α_{i1} -GDP- AlF_4^- -RGS9 at 20 mg ml⁻¹ with 1.5 μl of mother liquor (10.5% PEG 8000, 110 mM magnesium acetate, 50 mM Tris pH 8.5, 0.2% β -mercaptoethanol). Crystals were transferred into mother liquor plus 35% glycerol and flash frozen. Crystals belong to the space group P2₁2₁2₁ with two complexes per asymmetric unit. The α_{i1} -GDP- AlF_4^- -PDE γ -RGS9 complex was formed by incubating a 1:1 molar ratio of PDE γ and α_{i1} -GDP- AlF_4^- -RGS9 at 4 °C for 1 h. Heterotrimer crystals (native and RGS9_{Se-Met} derivatized) were obtained by mixing 1.5 μl of α_{i1} -GDP- AlF_4^- -PDE γ -RGS9 at 20 mg ml⁻¹ with 1.5 μl mother liquor (15.0% PEG 8000, 200 mM Tris pH 9.0, 0.2% β -mercaptoethanol, 1 mM (NH₄)₂WS₄). The cryopreservation protocol is analogous to that used for the heterodimer crystals. Crystals belong to the space group P2₁2₁2₁ and contain one α_{i1} -GDP- AlF_4^- -PDE γ -RGS9 complex and one α_{i1} -GDP- AlF_4^- -RGS9 complex per asymmetric unit.

Data collection, structure solution and refinement

Diffraction data were measured at the Argonne National Laboratory (ANL) Advanced Photon Source (APS) Structural Biology Center (SBC) 19-ID beamline, and processed and scaled using HKL2000 (ref. 28). Selenium MAD data were collected using inverse beam

geometry. Heavy-atom searches and phasing were performed using the CNS package²⁹. Selenium sites were found by using an automated Patterson heavy atom search method and peak anomalous wavelength data. MAD phasing used selenium coordinates and the high energy remote wavelength as reference. Phases were improved by solvent flipping and histogram matching, and where applicable, extended to the resolution of the native data set. Model building was done with O³⁰. Refinement was performed with the CNS package²⁹ and monitored with a random 10% of the data used for cross-validation. Model refinement employed the MLHL target function, torsion angle molecular dynamics simulated annealing, B-factor refinement, and rebuilding in O with σ_A -weighted difference Fourier maps. Three final rounds of refinement against an MLF target were performed for the α_{i1} -GDP- AlF_4^- -RGS9 structure using native data. Figures were prepared with MOLSCRIPT, RASTER3D, O and GRASP.

Received 11 September; accepted 20 November 2000.

- Angleton, J. K. & Wensel, T. G. A GTPase-accelerating factor for transducin, distinct from its effector cGMP phosphodiesterase, in rod outer segment membranes. *Neuron* **11**, 939–949 (1993).
- He, W., Cowan, C. W. & Wensel, T. G. RGS9, a GTPase accelerator for phototransduction. *Neuron* **20**, 95–102 (1998).
- Chen, C.-K. et al. Slowed recovery of rod photoresponse in mice lacking the GTPase accelerating protein RGS9-1. *Nature* **403**, 557–560 (2000).
- Skiba, N. P., Bae, H. & Hamm, H. E. Mapping of effector binding sites of transducin α -subunit using $\text{Go}/\text{G}\alpha_{i1}$ chimeras. *J. Biol. Chem.* **271**, 413–424 (1996).
- Skiba, N. P., Yang, C.-S., Huang, T., Bae, H. & Hamm, H. E. The α -helical domain of Go , determines specific interaction with regulator of G protein signaling 9. *J. Biol. Chem.* **274**, 8770–8778 (1999).
- Noel, J. P., Hamm, H. E. & Sigler, P. B. The 2.2 Å crystal structure of transducin- α complexed with GTP γ S. *Nature* **366**, 654–663 (1993).
- Lambright, D. G., Noel, J. P., Hamm, H. E. & Sigler, P. B. Structural determinants for activation of the α -subunit of a heterotrimeric G protein. *Nature* **369**, 621–628 (1994).
- Sondek, J., Lambright, D. G., Noel, J. P., Hamm, H. E. & Sigler, P. B. GTPase mechanism of G proteins from the 1.7-Å crystal structure of transducin α -GDP- AlF_4^- . *Nature* **372**, 276–279 (1994).
- Tesmer, J. J. G., Berman, D. M., Gilman, A. G. & Sprang, S. R. Structure of RGS4 bound to AlF_4^- -activated $\text{G}_{i\alpha 1}$: Stabilization of the transition state for GTP hydrolysis. *Cell* **89**, 251–261 (1997).

10. Spink, K. E., Polakis, P. & Weis, W. I. Structural basis of the Axin–adenomatous polyposis coli interaction. *EMBO J.* **19**, 2270–2279 (2000).
11. De Alba, E., De Vries, L., Farquhar, M. G. & Tjandra, N. Solution structure of human GAIP (Gα interacting protein): a regulator of G protein signaling. *J. Mol. Biol.* **291**, 927–939 (1999).
12. Tsang, S. H. *et al.* Role for the target enzyme in deactivation of photoreceptor G protein *in vivo*. *Science* **282**, 117–121 (1998).
13. Slepak, V. Z. *et al.* An effector site that stimulates G-protein GTPase in photoreceptors. *J. Biol. Chem.* **270**, 14319–14324 (1995).
14. Faurobert, E., Otto-Bruc, A., Chardin, P. & Chabre, M. Tryptophan W207 in transducin Tα is the fluorescence sensor of the G protein activation switch and is involved in the effector binding. *EMBO J.* **12**, 4191–4198 (1993).
15. Natochin, M., Granovsky, A. E. & Artemyev, N. O. Identification of effector residues on photoreceptor G protein, transducin. *J. Biol. Chem.* **273**, 21808–21815 (1998).
16. Muradov, K. G. & Artemyev, N. O. Loss of the effector function in a transducin-α mutant associated with Nougaret night blindness. *J. Biol. Chem.* **275**, 6969–6974 (2000).
17. Liu, Y., Arshavsky, V. Y. & Ruoho, A. E. Interaction sites of the COOH-terminal region of the γ subunit of cGMP phosphodiesterase with the GTP-bound α subunit of transducin. *J. Biol. Chem.* **271**, 26900–26907 (1996).
18. Artemyev, N. O., Rarick, H. M., Mills, J. S., Skiba, N. P. & Hamm, H. E. Sites of interaction between rod G-protein α-subunit and cGMP-phosphodiesterase γ-subunit: Implications for the phosphodiesterase activation mechanism. *J. Biol. Chem.* **267**, 25067–25072 (1992).
19. Tesmer, J. J. G., Sunahara, R. K., Gilman, A. G. & Sprang, S. R. Crystal structure of the catalytic domains of adenylyl cyclase in a complex with G_{sa}·GTPγS. *Science* **278**, 1907–1916 (1997).
20. Yan, S.-Z., Huang, Z.-H., Rao, V. D., Hurley, J. H. & Tang, W.-J. Three discrete regions of mammalian adenylyl cyclase form a site for G_{sa} activation. *J. Biol. Chem.* **272**, 18849–18854 (1997).
21. Lipkin, V. M., Dumler, I. L., Muradov, K. G., Artemyev, N. O. & Etingof, R. N. Active sites of the cyclic GMP phosphodiesterase γ-subunit of retinal rod outer segments. *FEBS Lett.* **234**, 287–290 (1988).
22. Brown, R. L. Functional regions of the inhibitory subunit of retinal rod cGMP phosphodiesterase identified by site-specific mutagenesis and fluorescence spectroscopy. *Biochemistry* **31**, 5918–5925 (1992).
23. Lambright, D. G. *et al.* The 2.0 Å crystal structure of a heterotrimeric G protein. *Nature* **379**, 311–319 (1996).
24. Natochin, M. & Artemyev, N. O. Substitution of transducin Ser²⁰² by Asp abolishes G-protein/RGS interaction. *J. Biol. Chem.* **273**, 4300–4303 (1998).
25. Nekrasova, E. R. *et al.* Activation of transducin guanosine triphosphatase by two proteins of the RGS family. *Biochemistry* **36**, 7638–7643 (1997).
26. McEntaffer, R. L., Natochin, M. & Artemyev, N. O. Modulation of transducin GTPase activity by chimeric RGS16 and RGS9 regulators of G protein signaling and the effector molecule. *Biochemistry* **38**, 4931–4937 (1999).
27. Sowa, M. E., He, W., Wensel, T. G. & Lichtarge, O. A regulator of G protein signaling interaction surface linked to effector specificity. *Proc. Natl Acad. Sci. USA* **97**, 1483–1488 (2000).
28. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
29. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
30. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron-density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).

Acknowledgements

We thank C. Berlot, Y. Korkhin, D. Lambright, L. Rice and J. Steitz for useful discussions; A. Brunger, P. Adams, R. Grosse-Kunstleve for help with CNS; and W. Minor and Z. Otwinowski for pre-release of HKL2000. We also thank staff of the SBC beamline 19-ID: R. Alkire, N. E.C. Duke, S. L. Ginell, K. Lazarski, S. Korolev, F. J. Rotella, R. Sanishvili, J. Lazarz, and A. Joachimiak. Use of the ANL SBC beamline at the APS was supported by the US Department of Energy, Office of Biological and Environmental Research. This work is supported in part by a NIH grant to P.B.S. M.A.K. is a HHMI Fellow of the LSRE. C.W.C. was supported by an NIH grant.

Correspondence and requests for materials should be addressed to K.C.S. (e-mail: kcs@mail.csb.yale.edu). Coordinates and structure factors are deposited in the Protein Data Bank under accession codes 1FQI, 1FQK and 1FQJ for RGS9, α_{u11}·GDP·AlF₄⁻·RGS9, and α_{u11}·GDP·AlF₄⁻·PDEγ-RGS9, respectively.



Cover illustration
by Jacey.

nature insight

Astrobiology

Commentators often bemoan what they see as the increasing tendency of research to be inward looking, self-serving and specialized. Happily, some things buck the trend. Astrobiology does this in dramatic fashion, and is the theme for the eclectic selection of review, progress and commentary that follows. Astrobiology is nothing less than the study of life in the Universe. It encompasses fields as diverse as geology, astronomy, evolutionary and developmental biology, human physiology and palaeontology.

A canvas so broad may lose its meaning — conversely, many researchers may be unaware that their narrow tasks could fall into the astrobiology net. To help define and explore the subject, a partnership between NASA and several universities has created the National Astrobiology Institute (see <http://nai.arc.nasa.gov>).

It would be easy to dismiss astrobiology as either a pointless fad or a new brand for goods long past their sell-by date (does anyone remember 'exobiology'?). But such condemnation misses the point. In an age of increased narrowness of research goals, it is invigorating to lift one's eyes to the stars and consider life in its broadest sense. One consequence of astrobiology will be to deepen our understanding of our own place in the Universe, our uniqueness and our potential.

On page 1080, Brian Aldiss suggests that astrobiology represents humanity's craving for otherness. Alone in the Universe, we demand others to share it, whether deities or aliens. Down to Earth, Euan Nisbet and Norm Sleep investigate the origin of life on page 1083. Did life originate on Earth, or elsewhere? As Lynn Rothschild and Rocco Mancinelli show on page 1092, life abounds even in the most uncongenial environments. "Normal is passé," is their slogan, "Extreme is chic." The human-centred prejudice for an oxygen-rich, desiccating environment is itself extreme, and could blind us to the possibility that life might be universally abundant. On page 1102, Sean B. Carroll considers what aspects of life on Earth are likely to be fundamental and thus shared with life elsewhere. He concludes that simple unicells might be common, but 'mega-organisms' rare. Undeterred, on page 1110 Tom Wilson considers the progress of SETI — the Search for Extraterrestrial Intelligence. But what of our own outward urges? On page 1115, Ronald J. White and Maurice Avernier look at the medical problems posed by the exposure of humans to the space environment. Implicit in our search for life elsewhere is the assumption that life is very much 'as we know it'. On page 1119, Ian Stewart and Jack Cohen explode this idea and ask whether, if extraterrestrials exist, we would be able to recognize them as such.

As this collection of articles shows, many ventures labelled as astrobiology are quixotic, even romantic, perhaps some way from the usual stuff of science. Disagreement abounds, and answers to many astrobiological problems may never be found. But that there is no universal truth is true for all science: the stimulation is in the quest, not in the finding. How much more valid is that statement when the quest encompasses the spatial and temporal breadth of the cosmos?

Henry Gee Senior Editor

Chris Surridge Insight Programme Editor
Liz Allen Publisher

Desperately seeking aliens

Brian W. Aldiss

39 St Andrews Road, Old Headington, Oxford OX3 9DL, UK

Belief that intelligent life is commonplace in the Universe was taken for granted by scholars and scientists until well into the nineteenth century. Space travel since the late 1950s reignited the debate, which even now attracts discussion by serious, professional scientists. And although statisticians might lobby that life must surely exist somewhere in the Universe, the evolution of what we perceive as 'intelligent life' seems utterly improbable — elsewhere as well as on Earth. Can we free ourselves of our animist fantasies and accept that all alien forms of intelligent life are, and always have been, imaginary?

It is easy to imagine the existence of life elsewhere in the Universe. The key word here is 'imagine' — the human mind has been populated with gods and demons since time immemorial, products of an apparently insatiable craving for the exotic. And still we yearn, our dreams turning from the supernatural and animist to the popular culture of such inventions as Mickey Mouse and Bugs Bunny, Klingons and Vulcans, and, of course, the *Alien*. The fruitfulness of our imagination is surprising in view of the fact that the Universe itself has offered no help: so far, our search for signs of alien life has drawn a blank. As far as we know, consciousness has dawned nowhere but on our home planet, Earth. I shall argue the case that — for the moment, at least — all other forms of intelligent life are imaginary, as they always have been.

The case that intelligent life is rare in the Universe is logical, yet it is hardly more than a century old, and showing signs of waning in the face of scientific initiatives such as the founding of the NASA Astrobiology Institute, whose aim is to explore the conditions for life on Earth and elsewhere, and even in the commission of this article for a *Nature* Insight entitled 'Astrobiology — Life in the Universe', in which the possibilities of life elsewhere in the Universe are discussed by serious, professional scientists. In the face of millennia of desperation to find aliens, recent scepticism, such as Brownlee and Ward's book *Rare Earth*¹, might be taken for a *fin-de-siècle* aberration.

A history of belief

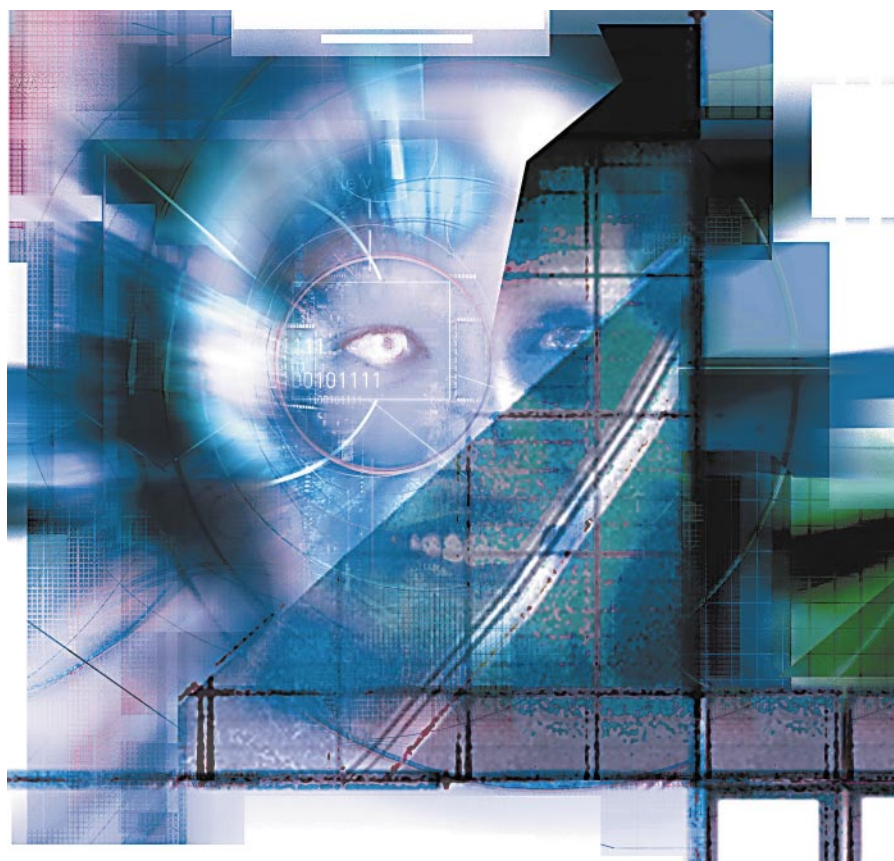
Serious speculation about life elsewhere was once commonplace. A few centuries ago, many scholars believed that intelligent life existed everywhere, and that an all-powerful God in his generosity had bestowed life on all the planets of the Solar System. This belief had firmest tenure on our neighbouring heavenly body, the Moon². We cannot tell how ancient this erroneous belief may be, but the first story to be set on the Moon is generally agreed to have been written in the second century AD by Lucian of Samosata, whose *True History* is a satire on travel writing.

Lucian's travellers are carried by a waterspout in a Greek ship to the Moon. There they discover that the King of the Moon and the King of the Sun are at war over the issue of the colonization of Jupiter. Fantastic monsters are employed in battles on both sides. Such adventures have always been popular, at least from recent centuries onwards. One authority, Philip B. Gove, lists 215 books describing voyages to the Moon published in the eighteenth century alone³. Modes of transport have varied, from angels to migratory geese.

Science has always provided the most potent fuel for the imagination. Space fiction took off after Galileo published *The Starry Messenger* in 1610, conveying vividly the excitement of the moment when a man first looked through a telescope into space. Not only was the Moon no perfect sphere, as had been always thought, but was "just like the surface of the Earth itself, varied everywhere by mountains and valleys". Following his description of the Moon, Galileo went on to reveal his discovery that "there are not only three, but four, erratic sidereal bodies performing their revolutions round Jupiter". This observation of the four main jovian satellites overturned the old Aristotelian thinking, which had set the Earth at the centre of the Universe. Galileo's name became celebrated beyond his native Italy. No longer was it possible for informed people to believe that the Sun went round the Earth. Henceforth, the heliocentric version of our Solar System would prevail, and bring forth many celestial tales — generally satires or utopias. The telescope fathered both astronomy and fantasy. Just one example was *Man in the Moone* (1638) by the learned Bishop Francis Godwin of Hereford, which remained in print for more than two centuries and was much translated. Possibly because the bishop considered his book went against the teachings of the Church, it had to await publication until after his death.

That life in the Universe was, well, universal was taken for granted in the scientific sphere until well into the nineteenth century. William Whewell, the scientist who famously coined the word 'scientist', found it necessary to dispute the belief in universal life. His book *Of the Plurality of Worlds* was published anonymously in 1855. Not that Whewell's views did anything to stem the tide of aliens in fiction. Since the days of H. G. Wells, when cars replaced horses, writers have propagated aliens with increasing assurance. If aliens do not exist, it seems necessary to invent them. It is a nice irony of modern life that the prospects of finding real-life aliens have dimmed just as the 'realism' of fictional aliens has waxed. Perhaps the two are connected — and yet the pendulum could be swinging back sharply.

By the late 1950s, the idea of intelligent life on Mars or any other planet was unfashionable enough to be the subject of derision. The tide turned just two weeks after the Astronomer Royal, Sir Harold Spencer Jones, announced in 1957 that space travel was bunk — when the Soviet Union sent up the first Sputnik. (Jones later compounded his error by saying that he was talking about science fiction.) Once it was generally realized that large objects could travel through space, propelled by rocket motors, the gates were open for speculation about visits and visitations to and from Earth. It was a technological dream. From then



onwards, it seemed that most people in the West believed — as had the ancients — that all about us were unseen planets of stars abounding with life. For all Whewell's work, the notion of plurality of interplanetary life had returned. By the early 1960s, unthinking scepticism had turned to unthinking belief.

Earth's neighbours and beyond

Nothing except statistics supports the idea that life (or at least intelligent life) exists anywhere else but the Earth. The evidence in our own Solar System is decisively negative. The Moon as an abode of life was ruled out when it was discovered that it had no atmosphere. Elimination next for our shrouded neighbour Venus, of which the Swedish astronomer, Svante Arrhenius, deduced in 1917 that "everything on Venus is dripping wet"⁴. The surface, according to Arrhenius, was covered by swamps, in which low forms of life existed: "the organisms are nearly of the same kind all over the planet". (In a forgotten novel of 1956, *Escape to Venus*, S. Makepeace Lott is nearer the mark, speaking of "the battering of the gas storms which flung the suspended dust particles across the face of the planet at several hundred kilometres an hour".) With a mean surface temperature of 740 K, Venus is an unlikely abode of life.

So to Mars, the planet on which most expectations of finding life were pinned. In 1909, astronomer Percival Lowell — self-delusive finder of martian canals —

published the well-reasoned *Mars as the Abode of Life*. It must have seemed reasonable at that period to believe in life on our dry neighbouring planet, when the previous century had uncovered evidence of a staggering abundance of life, never previously dreamed of and flourishing over millions of years, in the strata of terrestrial rock. If a monstrous fossil reptile in the ancient sandstone, why not a little green man on Mars?

But no. Since Lowell's day, Mariners and Vikings have called on Mars. Dust and rocks are all they have found. Mars is a bleak, stony place: dry, with only the thinnest of atmospheres. Viking revealed the martian surface as a highly inhospitable environment for life. The finding of microscopic impressions in a meteorite, believed to be of martian origin, and which might, in some circumstances, have been fossils, has been controversial.

Venus, Earth and Mars lie in the Sun's 'comfort zone'. Beyond Mars stretches a gulf of space, with the gas giants beyond it — surely, there can be no hope for life out there? But the Galileo spacecraft has produced strong evidence that beneath the icy and broken surface of Europa, one of the four galilean Moons of Jupiter, lies an ocean⁵, warmed by the gravitational pull of Jupiter. What might we anticipate there? Intelligent shrimps? Intellectual fish? We can but hope — but there is still a line to be drawn between hope and conviction.

And beyond the Solar System? Our Galaxy contains approximately 200 billion

stars. Surely some of them must have planets that sustain life? It is not an unreasonable conjecture, given the numbers. Although we have no evidence that any of the now several dozen known extrasolar planetary systems⁶ have suitable conditions for life of the kind we might recognize as such, the numbers could give us hope.

Improbable evolution of intelligence

But statistical casuistry works both ways, as is shown by the improbability of intelligent life appearing on the only planet we know well — the Earth. Although life appeared on Earth at least 3.8 billion years ago, not long after the planet itself formed (see the review in this issue by Nisbet and Sleep, pages 1083–1091), it took another 3.2 billion years before the appearance of complex, multicellular life forms large enough to be viewed without a microscope. Intelligence (as we perhaps mistakenly understand it) has developed only in the past few tens of thousands of years. According to Ward and Brownlee⁷, microbial life in our Galaxy might be common, but complex, multicellular life will be extremely rare.

Each of the steps — between the appearance of life and the evolution of intelligence — reveals its complexity, helped on or deterred by coincidences and catastrophes. Moreover, there might have been only one time propitious for creating the rudiments of life: later might have been too late. Given its evolution through a number of precarious episodes, we perceive that 'intelligent life' is an uncharacteristic effect, not merely in our own Solar System but more universally. In fact, it seems utterly improbable — elsewhere as well as here.

This knowledge has not deterred serious-minded people from attempting to make contact with intelligences elsewhere in the Galaxy⁷. The Search for Extraterrestrial Intelligence (SETI) programme was set up in the 1960s, although so far no one or nothing has answered its signals (see the review in this issue by Wilson, pages 1110–1114). Nor have we heard any signals from elsewhere.

A challenge to the consensus of universal biological ubiquity was presented in 1986 by John D. Barrow and Frank J. Tipler in *The Anthropic Cosmological Principle*⁸, a powerful sequel to Whewell's argument. Using many disciplines, the authors argue that, by an element of design, ours is the only planet that houses cognate beings. Their argument is complex, encompassing the stability of stars and the eccentricities of water, on which life and its origins depend heavily. In sum, it leaves human cognition with a large responsibility for acting as the consciousness of the Universe.

C. O. Lovejoy is quoted as saying: "Man is not only a unique animal, but the end product of a completely unique evolutionary pathway, the elements of which are

traceable at least to the beginnings of the Cenozoic.”⁹ This pathway is defined by the evolutionary biologist Ernst Mayr. Speaking of the principal divisions (or phyla) in the animal kingdom, he says that the kingdom “consists of about 25 major branches... Only one of them developed real intelligence, the chordates. There are numerous classes in the chordates, I would guess more than 50 of them, but only one of them (the mammals) developed real intelligence, as in Man. The mammals consist of 20-odd orders. Only one of them, the primates, acquiring intelligence, and among the well over 100 species of primates only one, Man, has the kind of intelligence that would permit the development of advanced technology... An evolution of intelligence is not probable.”¹⁰

The blessing of science

We understand that optimism and imagination help to propel science. Nevertheless, we are entitled to ask whether assumptions about alien life are unscientific. Aliens are the staple diet of modern entertainment, but these are, in the main, contemporary fairy stories, and none the worse for that. However, their relationship with real science is ambiguous. Imaginary aliens are many and diverse, but provide little help in any current comprehension of understanding the Universe: rather than assisting us, aliens impede understanding. Their air of seeming rationality, of being the product of scientific thinking, is spurious. Where, then, do aliens originate, and how has our desperate search for aliens come to find itself on any serious scientific agenda?

An intimacy with the non-human is a fundamental human trait. A vast population of ghosts, ghouls and other mythical creatures has accompanied humankind through the ages, haunting its woods, houses and graveyards. Among their attractions is that they are free of the physical laws that govern humans. In particular, they are at least partly immune to gravity and

death (a tradition continued among mythical cartoon creatures such as Tom and Jerry).

Above these minions, as religion outranks superstition, are assembled an even more formidable array of fictitious beings, the gods and goddesses of our inner world. What a collection they are! Belief in them beggars belief: adorned with snakes and skulls, they arrive to impose restrictive laws for human conduct, laws that frequently include whom we should or should not sleep with, and the preservation of life and the sacrifice of it. Coming from a generation which listened to damnation preached every Sunday, brought up to believe in a cloudy Heaven and the fiery torment of Hell (ruled over by a horned and unpleasant Satan), I now recoil from the cruelty of the pulpit, and can but marvel at the entire range of weird deities.

We do not believe in fairies any more, nor do we find it necessary to blaspheme against Baal. But it seems that we are born animists. Parents heap a variety of totemistic animals on their children: *Tyrannosaurus rex* is to be found sharing the cot with Winnie the Pooh. As children talk to their stuffed toys, so adults talk to their pets and pray to one or more members of an invented pantheon.

The latest manifestation of this creaking floorboard in the brain, the alien arriving here from outer space, is the most interesting. Such an event could conceivably happen, and may be regarded indulgently as more supposition than superstition. Much work has been done to render this magical visit plausible. In the 1960s television drama *A for Andromeda*, written by John Elliott and Fred Hoyle, radio signals emanating from the Andromeda Galaxy are picked up by the then new radio telescope at Jodrell Bank, near Manchester, United Kingdom. The signals include directions for the construction of a computer. This computer enables the scientists to build a beautiful alien woman (the first appearance on our screens of Julie Christie). *A for Andromeda*, broadcast hardly

an eyeblink beyond the launch of the first Sputnik, marks the emergence of alien life from fantasy into cool scientific reality, given the blessing of a computer. Science fiction infiltrates science itself.

Julie Christie, if memory serves, was gracious and a source of wisdom in her alien avatar. Sometimes, aliens arrive to save us from our own follies. More frequently, they come to invade and destroy us. Such thinking forms a continuity with our ancient dreads of demons, ever hostile to human life.

Let us suppose that aliens are, as I have suggested, merely the latest example of a form of animism at work: or possibly the immature echoes of our own selves, free of time and gravity. So let us suppose further that no one will ever visit or call — because no one is there to call. We, the entire riotous biomass of Earth, are alone on our small planet.

The implications of such a situation are formidable. Scientifically and philosophically, a change of attitude would be demanded. In *A Defence of Poetry* (1821), Shelley states that “man, having enslaved the elements, remains himself a slave”. Could we but free ourselves from those atavistic fancies here enumerated, humankind might consider it not impossible that we should go into the Galaxy with the intention of becoming its consciousness. □

1. Ward, P. D. & Brownlee, D. *Rare Earth: Why Complex Life is Uncommon in the Universe* (Copernicus Books, New York, 2000).
2. Nicolson, M. H. *Voyages to the Moon* (Macmillan, New York, 1960).
3. Gove, P. B. *The Imaginary Voyage in Prose Fiction* (Columbia Univ. Press, New York, 1941).
4. Arrhenius, S. *The Destinies of the Stars* (transl. Fries, J. E.) (Putnam's, New York, 1918).
5. Carr, M. H. Evidence for a subsurface ocean on Europa, *Nature* **391**, 363–365 (1998).
6. The search for extrasolar planets <<http://exoplanets.org>>.
7. Heidmann, J. *Extraterrestrial Intelligence* (Cambridge Univ. Press, Cambridge, 1995).
8. Barrow, J. D. & Tipler, F. J. *The Anthropic Cosmological Principle* (Clarendon Press, Oxford, 1986).
9. Lovejoy, C. O. in *Life in the Universe* (ed. Billingham, J.) (MIT Press, Cambridge, MA, 1981).
10. Mayr, E. Letter to Frank Tipler, December 1982. Quoted in Barrow, J. D. & Tipler, F. J. *The Anthropic Cosmological Principle* (Clarendon Press, Oxford, 1986).

The habitat and nature of early life

E. G. Nisbet* & N. H. Sleep†

*Department of Geology, Royal Holloway, University of London, Egham TW20 0EX, UK

†Department of Geophysics, Stanford University, Stanford, California 94305, USA

Earth is over 4,500 million years old. Massive bombardment of the planet took place for the first 500–700 million years, and the largest impacts would have been capable of sterilizing the planet. Probably until 4,000 million years ago or later, occasional impacts might have heated the ocean over 100 °C. Life on Earth dates from before about 3,800 million years ago, and is likely to have gone through one or more hot-ocean ‘bottlenecks’. Only hyperthermophiles (organisms optimally living in water at 80–110 °C) would have survived. It is possible that early life diversified near hydrothermal vents, but hypotheses that life first occupied other pre-bottleneck habitats are tenable (including transfer from Mars on ejecta from impacts there). Early hyperthermophile life, probably near hydrothermal systems, may have been non-photosynthetic, and many housekeeping proteins and biochemical processes may have an original hydrothermal heritage. The development of anoxygenic and then oxygenic photosynthesis would have allowed life to escape the hydrothermal setting. By about 3,500 million years ago, most of the principal biochemical pathways that sustain the modern biosphere had evolved, and were global in scope.

What is life? Natural science has never found a satisfactory definition. Cardinal Newman¹, following Thomas Scott, lived by the parable “Growth is life”. This is not the answer to the puzzle, but it reduces the problem by providing a useful working tool to the geologist. Life can be recognized by its deeds — life is disequilibrium, leaving behind the signatures of disequilibrium such as fractionated isotopes or complex molecules. It is more besides, but the larger question ‘what is life?’ is perhaps beyond natural science. Continuum exists between chemistry, autocatalysis and what every one would agree is life. But defining the point at which autocatalysis becomes life is like searching for the world’s smallest giant.

The prebiotic Hadean environment

The Solar System began after one or more local supernova explosions about 4,600 million years (4.6 Gyr) ago. In one widely accepted scenario of the later stages of accretion of the Solar System it is thought that there were 500 or so planetesimals, bodies about the size of the Moon, in the region now occupied by the inner planets^{2,3}. Collision between planetesimals produced the inner planets. Venus, Earth and Mars all received inventories of water vapour and carbon, perhaps with early oceans on all three. But other models are also possible⁴. The fate of the volatile inventory⁵ in each planet was completely different: Venus is dry, with a surface now at around 500 °C under 90 bar (9×10^6 Pa) of carbon dioxide (CO₂). Mars is in permafrost. Earth has approximately the same external inventory of CO₂ as Venus, and both planets radiate heat to space at very similar ‘effective’ temperatures (in some senses Earth is hotter), but for us the CO₂ is tied up in carbonate minerals (for example, limestone). The blanket is less and so the oceans can exist.

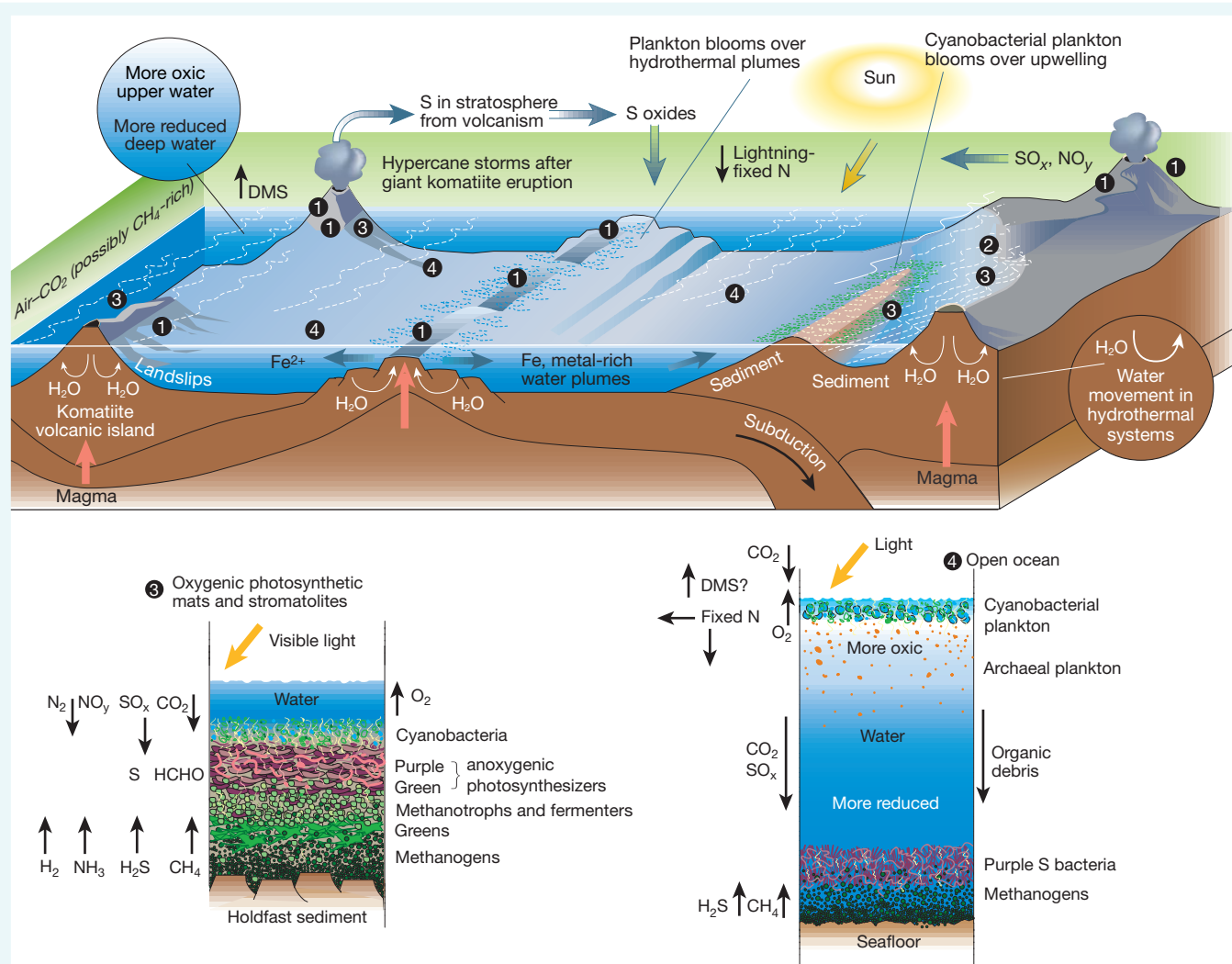
Water is a strong greenhouse gas and, at some stages early in the history of Venus and Earth, water vapour was probably present high in the atmosphere. Such water vapour would have been photolysed into hydrogen and oxygen, and the hydrogen present in the upper atmosphere would have been lost rapidly to space. Deuterium would have been lost also, but being more massive, would have been lost more slowly. In comparison to the D:H ratio of hydrogen that is thought to have been originally in the planetary mix, the residual hydrogen in Venus’s

atmosphere has a strong deuterium enrichment. The simplest explanation is that Venus lost its water early in its history when a runaway greenhouse developed. In this model, initially Venus had oceans and a warm (>75 °C) surface; water was partitioned into the high atmosphere, photolysed, and hydrogen was lost and the planet dehydrated⁶, leaving a more oxidized planet. Alternately, if Venus has or had a molten magma ocean in its mantle it may there too have sequestered ‘light’ hydrogen as OH. Mantle minerals are typically ‘light’, or depleted in deuterium relative to sea water⁷.

Modern Earth does not lose significant amounts of water to space. Today, the top of Earth’s troposphere is cold and water falls back to the surface: the stratosphere is extremely dry, and relatively little hydrogen reaches the top of the atmosphere. It is possible that on the early Earth, as probably on early Venus, there was a substantial hydrogen loss to space, but to a lesser extent, losing only up to a little over a third of the ocean⁸. If so, the residual oxygen (up to 100 bar) would have meant that the outer part of the young Earth was significantly enriched in oxidant compared to the interior. The noble gases of Earth’s atmosphere have a complex history⁹, in part recording primordial accretion in planetesimals, but perhaps also recording the early onset of subduction.

On Mars, residual water is today present as subsurface brine aquifers subject to rare break-out floods; there is evidence for earlier events when water was free on the surface¹⁰. Mars also had early volcanic activity. With water and volcanoes, early Mars may have been an eminently habitable place, perhaps more so than Earth.

Earth is unique among the inner planets in possessing a disproportionate Moon. Of the explanations for the origin of the Moon, the most persuasive is that proto-Earth was struck, around 4.5 Gyr, by another inner planet about the size of Mars or even larger¹¹. This impact spun and tilted the Earth¹². The heat of the impact would have been enough to melt the Earth, even if it were not already molten from infall heat and the heat from radioactive decay. The impact (the ‘big splat’) would have ejected enormous amounts of molten mantle into orbit, some of which coalesced to form the Moon. The consequences of the impact are wide. The spin and tilt of the Earth have evolved from that event, dissipated by tides, to give us the present day–night cycle and the seasons, both crucial in distributing heat around the planet and making the Earth habitable by



biology. The heat that powers modern plate tectonics and volcanism originates in approximately equal measures from the primordial impact heat of accretion and early radioactivity, and from the long-term decay of uranium, thorium and potassium. Were it not for the continual renewal of the erosion cycle driven by this heat, and the supply of new volcanic cations and reduction power, the productivity of the biosphere would have slowly died from shortage of essential elements. Instead, nature remains unimpaired by time¹³.

Immediately after the Moon-forming impact, a rock vapour atmosphere would have formed. As this cooled it would have formed an optically thick layer of dust in the high atmosphere. After about 2,000 years, a rind would have begun to form on the magma ocean of the mantle, and for perhaps 2 million years surface temperatures lingered near 100 °C, with a steam greenhouse, before beginning to cool. Bombardment on a lesser scale was ongoing, but gradually decaying. Until about 3.8 Gyr, the Earth would have suffered frequent massive meteorite impacts, some sufficiently large to heat the oceans to >110 °C, or even to the ~350 °C needed to convert the whole ocean to steam. Impacting bodies of a diameter of 500 km or more would have been capable of vaporizing the ocean; bodies of 200 km could have heated it above 100 °C. Impacts would have ejected huge quantities of debris, made of basalt or komatiite (a magnesium-rich lava). Planets are active, and the Earth's ocean is not simply an unchanging puddle. Seawater chemistry would have been controlled by volcanism and by reaction with this debris, such that it is improbable that a long-lived global 'primeval soup' could have collected from impacts of organic-rich meteorites and comets.

The oldest minerals known on Earth are 4.4-Gyr zircons from Western Australia¹⁴. These zircons are recycled — they now occur in a younger rock that eroded from a Hadean parent (see Box 1 for the aeons). The oldest rock (coherent assemblage of minerals) is the 4-Gyr Acasta gneiss from northwest Canada¹⁵. That it is a gneiss suggests that silica-rich rocks and possibly subduction existed at the close of the Hadean and the start of the Archaean. Perhaps water oceans existed, which were needed to provide cool, hydrated lithospheric plate for subduction and hence to create continental rocks^{14,16}.

The meteorite bombardment, heaviest until about 3.8 Gyr but to an extent continuing to the present (it remains a danger), was capable initially of ejecting matter into orbit from any of the inner planets. In rare cases, ejecta could have suffered very low stresses. Life, once begun, spreads rapidly into any available habitat. Thus if life were present on any inner planet with liquid water, it could have spread quickly across that planet, such that cells were present everywhere that conditions were suitable. Thereafter, each significant impact event producing huge numbers of ejecta would have been likely to have thrown into space a few rocks that carried undamaged cells. Some of these ejecta would have crossed to other inner planets: material from Mars still falls on Earth¹⁷. Had life existed on Mars, these ejecta could have carried one or more cells, quick frozen in space and held safe from ultraviolet (UV) light in rock crevices. Even if they survived ejection, nearly all cells on a rock would have perished in space, but it takes only one cell to infect a planet.

This opens the question: on which planet did life begin? Mercury is improbable, being sunward of the habitable zone and too extreme¹⁸. Venus, being large, would have had less planet-leaving ejecta, and events

② Anoxygenic photosynthetic mats

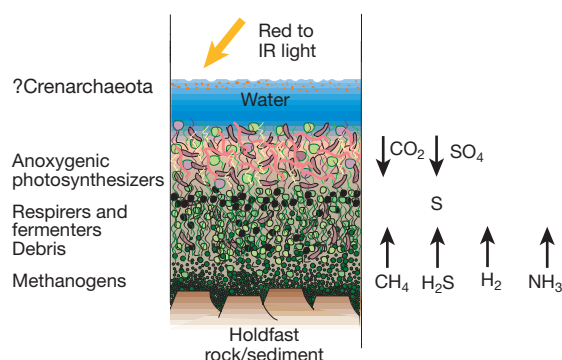
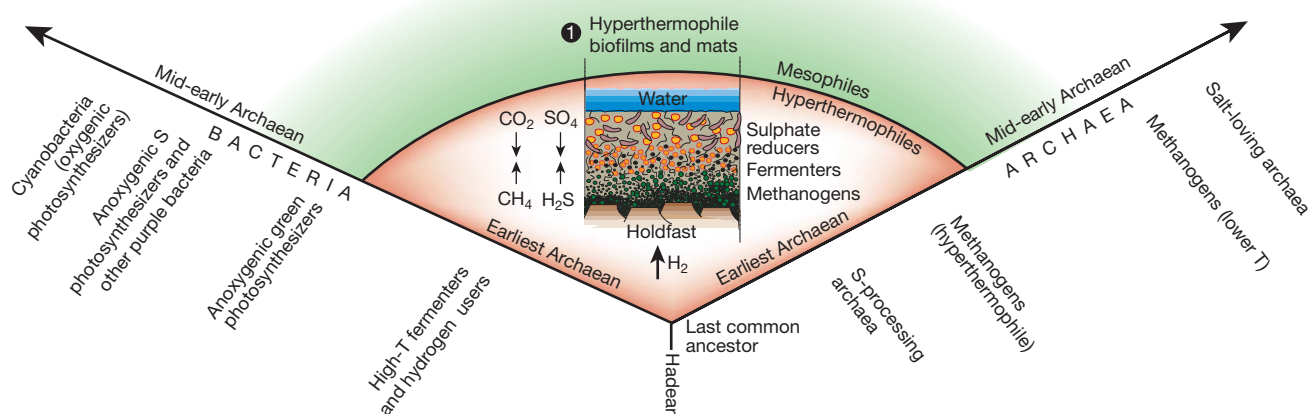


Figure 1 Late-Archaeon biosphere — the living communities and their chemical products. The upper part of the left panel shows a model of possible habitats of microbial communities. Field and isotopic evidence exists for many of these settings, but the presence of plankton is inferred from sediment record and molecular interpretation, and the mid-ocean ridge community is inferred. (Figure not to scale.) Microbial mat communities are illustrated in the lower part of the left panel and the right panel. Columns show possible mat communities and biofilms (numbers refer to typical settings in the habitat model). Evolutionary heritage follows standard model.



capable of ejecting undamaged cells into space would have been rare. However, a Perelandrian¹⁹ origin of life in early Cytherean oceans is not impossible. Mars is more suitable: it is a smaller planet and could have provided a suitable nurturing place given its possible early shallow seas under a CO₂-rich air, into which volcanoes would have been erupting²⁰. Being small in comparison to Venus, major impacts would have ejected rocks relatively frequently that were capable of reaching Earth. Impact sterilization would have been a significant danger to life on any inner planet. Overall, early Mars may have been safer than the early Earth, and Mars was possibly habitable in the Hadean²¹. The 'giant impactor' that hit Earth is also a possible although improbable birthplace, with a small chance that cells survived the impact by being ejected into orbit until the Earth's surface cooled or were transferred to Mars or Venus by the collision. Exchange between the inner planets is ~10⁶ times easier than exchange between the inner planets and outer planet satellites. Transfer from shallow interiors of large asteroids to inner planets is possible, but soft landing of debris on asteroids is unlikely. Finally, it is worth noting that if life did originate on Earth, rather than being carried here from Mars or another planet, it is possible that a meteorite impact could have ejected terrestrial living cells to infect Mars or Venus, despite the difficulty of leaving our relatively large planet.

The 'faint young Sun' puzzle

If the modern Earth were an airless, rapidly rotating planet of its present colour, its surface temperature — the 'effective temperature' T_e — would be 255 K (the effective temperature of modern Venus is similar)⁴. Because of the H₂O–CO₂ greenhouse, the actual surface is 33 K warmer

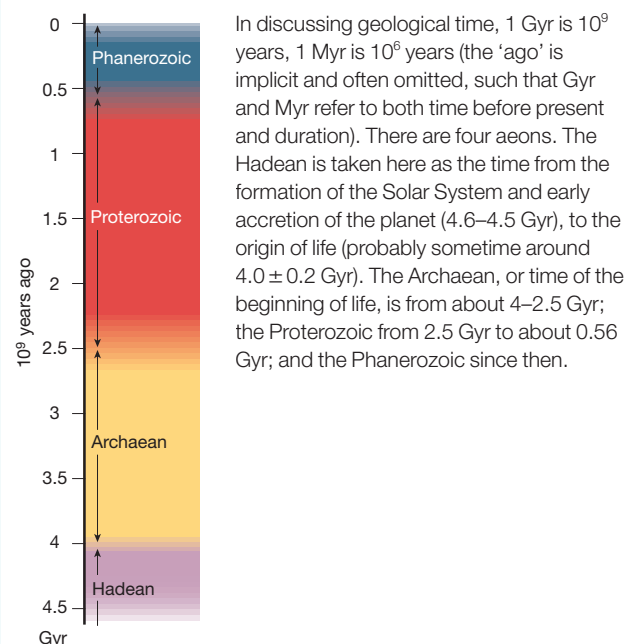
and thus sustains liquid water (which in turn allows abundant water vapour in the air, the main component of the greenhouse). The early Sun was fainter by a factor of roughly a quarter to a third²², and if such a sun shone on the present-day Earth there would be global glaciation. On the early Earth, therefore, the possibilities are either that there was indeed global glaciation, or that warmth was sustained by a massive global CO₂ greenhouse: a 10–100-bar CO₂–CO atmosphere over liquid ocean, with the CO₂ content of the air declining as the Sun warmed.

A CO₂ greenhouse would imply CO₂-rich oceans, which would react with basalt to form carbonates, especially with the huge supply of basalt fragments from impact ejecta. A CO₂ greenhouse would also have evolved to lower temperatures until volcanic outgassing of CO₂ matched the carbonate precipitation. This would take millions of years, during which the water temperature would linger near 100 °C. Thereafter, surface temperature may have been cool, even glacial. It is possible that the Hadean between 4.4 and 4.0 Gyr was mostly a Norse Ice-Hades, with intervals of Inferno after major impacts occurred when the Earth's surface was a lake of fire and brimstone. A glacial Earth in the Hadean would not lose much hydrogen to space, as so little water would be present in the high atmosphere. On Venus, being closer to the Sun, the slightly warmer temperatures may have allowed a moist greenhouse to be sustained in the long gaps between impacts, and the planet was consequently dehydrated.

Alternative hypotheses can be proposed. Ammonia and methane are strong greenhouse gases, and if their mixing ratio were high enough, they could have sustained liquid oceans. However, ammonia photodissociates quickly in solar UV. It is possible that methane in the

Box 1

Geological timescales



atmosphere could have produced a high-altitude organic smog that absorbed UV, allowing ammonia to exist below^{22,23}. In such circumstances, a liquid ocean could exist. However, for a methane smog to screen the planet, the $\text{CH}_4\text{:CO}_2$ ratio of the air must have been high (above unity²⁴). Methane can be produced after impact fireballs²⁵, but a long-lived methane atmosphere is perhaps unlikely on a prebiotic planet if surplus oxidation power comes from hydrogen loss⁸, but might have been possible later after methanogens (methane-generating organisms) had evolved. Reaction between a hydrous mantle magma shell and a reduced core²⁶ could also have provided oxidation power, through mantle-derived volcanic output.

More generally, a strongly reduced prebiotic atmosphere rich in methane and ammonia seems unlikely, given the probable oxidation state of lavas, and of the rock debris ejected from the mantle after major impacts, especially after differentiation of the iron-rich core. Perhaps abiotic Earth was indeed glacial, with occasional meteor-impact melting events, and pools of water around volcanoes. The ocean may have been covered with sea ice, with some open leads and thin ice where ablation occurred. Possibly on Venus a much warmer climate prevailed with early liquid oceans and a moist CO_2 greenhouse, but at the price of losing hydrogen to space. On Mars, further away but with a gentler impact history, a CO_2 greenhouse (or even a $\text{CH}_4\text{--NH}_3$ atmosphere) may have been sustainable for some time.

Summary of the geological evidence for early life on Earth

Earlier Archaean (4–3.6 Gyr)

Geological evidence^{27–29} shows that it is certain that life has been present on Earth for at least 3.5 Gyr, and it is probable that life began before 3.8 Gyr. Early evidence for life comes from southwest Greenland, especially the Isua belt which contains rocks up to 3.8 Gyr old that show clear evidence of deposition on the planet's surface, including what may be the oldest known water-lain sediment³⁰. There has long been a suspicion that inorganic carbonate in the rocks shows a ^{13}C 'heavy' enrichment that is the necessary counterpart of the light carbon extracted by the biosphere^{31,32}. But dating is controversial, and the carbonate is not necessarily the age of the host rock. In rocks from Isua and also from Akilia island, both in southwest Greenland, isotopically 'light' (that is, probably biologically reduced) carbon occurs in carbonaceous

inclusions in apatite (calcium phosphate)³³. A different line of evidence comes from sedimentary rocks in Isua that contain minute graphite globules with a ^{13}C content of about 19‰ (ref. 34). These may have been derived from early Archaean plankton. Controversy continues.

Mid-Archaean successions (3.6–3.3 Gyr)

In rocks around 3.5 Gyr old there have been many claims of fossil microbial biofilms and stromatolites (organosedimentary structures produced by microbial trapping, binding and precipitation, usually but not always photosynthetic)³⁵. Although some of these features can be interpreted equally as non-biological in origin³⁶, others possess all the diagnostic features of biogenic structures³⁷. Carbonate that is isotopically similar to modern carbonate is known throughout the past 3.5 Gyr, as is reduced carbon that shows isotopic fractionation that can only be explained as biological^{31,32}. The most obvious explanation is that throughout this time large-scale photosynthetic carbon fractionation (that is, oxygenic photosynthesis) by the enzyme ribulose-1,5-bisphosphate carboxylase-oxygenase or 'Rubisco' had operated — on a global scale. This removed about one-fifth of carbon released from the Earth's interior as 'light' biological carbon, leaving the counterpart 'heavier' four-fifths to precipitate as carbonate.

The evidence is not just isotopic. In rocks of about 3.3–2.5 Gyr in the Barberton Mountain Land in South Africa and the Pilbara in Western Australia there are carbonaceous microstructures that may have been microbial in origin^{38–40}. Filamentous microfossils are known in a 3.2-Gyr volcanogenic massive sulphide (deep-water) deposit in Western Australia⁴¹, implying that life existed on Archaean mid-ocean ridges⁴². There is thus reasonable ground to suppose microbial life was widespread in the mid-Archaean, probably present both on coastal fringes and as photosynthetic plankton in deeper water, as well as in both shallow and deep hydrothermal habitats.

Late-Archaean successions (3.0–2.5 Gyr)

In late-Archaean rocks there is abundant evidence for life^{28,42}. Stromatolites are well developed at Steep Rock, Ontario⁴³, and in the Pongola Supergroup, South Africa⁴⁴, which are both nearly 3.0 Gyr old. Well developed stromatolites from 2.7-Gyr successions are found in many places^{28,45,46}. Figure 1 shows various habitats of microbial mats and biofilms in the late-Archaean biosphere.

Textural evidence for late-Archaean bacteria depends on the assumption that similarity to modern microbial sedimentary structures implies similar ancient biota. Poorly preserved microfossils³⁹ and indirect geochemical arguments⁴⁷ supported the case, but it remained circumstantial. More direct evidence has now come from molecular fossils — biological lipids are preserved in 2.7-Gyr rocks from the Pilbara, Western Australia⁴⁸. Hydrocarbon biomarkers, including 2α -methylhopanes which are characteristic of cyanobacteria⁴⁹, imply oxygenic photosynthesis was occurring. Steranes, derived from chemicals characteristic of eukaryotes, are also present. Whether the sulphur cycle was similarly modern in aspect is at present controversial⁵⁰.

The universal ancestor and the last common ancestor

Insight into the descent of life has come from molecular palaeontology⁵¹. The 'standard model' of microbial descent^{52,53} is based on small-subunit ribosomal RNA. One hypothesis^{54,55} is of an early population of replicating organisms of uncomplicated design, possessing simple modular structures and functions, and mutually exchanging genes — the 'universal ancestor' was not one cell but a community sharing information. As evolution selected proteins to become more specific and efficient, genes became less exchangeable and divergence crystallized. From this 'universal ancestor' the 'standard' view^{51,56} is that the two domains Bacteria and Archaea arose, and that later on, further along the 'universal phylogenetic tree', symbiosis produced the domain Eucarya^{57,58}.

The standard view⁵² clearly implies (but does not prove) that in the early microbial community in which the last common ancestor lived, life was hot and chemotrophic⁵⁹ — the 'hyperthermophile Eden' hypothesis — although evidence for a hyperthermophile

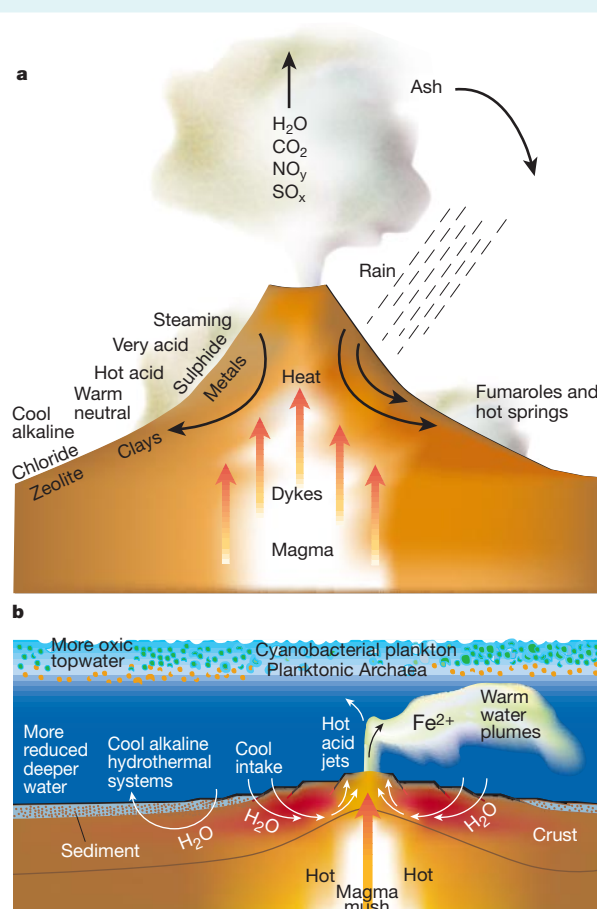
Box 2

Hydrothermal systems

Hydrothermal systems occur wherever magma reaches the surface. Water is heated by the magma, becomes less dense, and rises to be replaced by incoming cool fluid. As it moves it interacts chemically with the rock matrix, leaching metals. When it emerges and suddenly cools, these are redeposited. Around volcanoes on land, rainwater-fed systems form hot pools and fumaroles of varying pH. At mid-ocean ridges, in water 2.5–3 km deep, vigorous circulation includes ‘black smokers’ emitting jets of very acid water at temperatures up to 400 °C, which are crucial in ocean chemistry, especially metal supply and pH control^{106,107}. Many minerals offer internal surfaces that are organophilic and catalytic¹⁰⁸.

Modern subaerial hydrothermal pools are widely colonized by hyperthermophile bacteria and archaea. Subaerial systems are very diverse, with a wide range of pH possible. Near-surface magma can heat steam to >600 °C, with highly acidic, vigorously boiling hot springs. Fluids heated by basaltic magmas degassing at depth can form near neutral to alkaline springs. If the country rock is ultramafic (magnesium-rich), very alkaline systems can occur. Komatiite shield volcanoes may have produced many alkaline systems. At high levels in the volcano both fluid and vapour phases can occur. Country rock is altered to clay, and sulphide deposition (usually iron sulphide) is widespread. Komatiites host nickel sulphide deposits²⁸.

Submarine systems¹⁰⁷ include diffuse vents as well as black smokers, which emit iron and manganese oxides. Typically, hydrothermal fluids that have interacted with magma in some way are more reduced than overlying sea water. Sulphur has a key role. Some is volcanic, but in modern systems much of the sulphur is from seawater sulphate (which derives from sulphur gases via photosynthetic oxidation). The sulphate is reduced bacterially (for example, against organic matter, or against more reduced chemical species) as the water enters the hydrothermal circulation, further reduced in the circulation system, and then reoxidized by bacteria as the water leaves vents and mixes with ambient oxidized water. Similarly, nitrate is reduced, such that ammonium minerals are found in deposits. Before photosynthesis, the supply of sulphate and nitrate (that is, oxidized species of sulphur and nitrogen) to oceanic water was probably far less than today, coming from, respectively, oxidation of sulphur gases by OH in the upper air, and lightning fixation of nitrogen, and returned by volcanism.



Box 2 Figure Hydrothermal systems. **a**, On land, around a volcano. **b**, On seafloor, at a mid-ocean ridge. (Not to scale.)

ancestry has been challenged⁶⁰. Several geological settings could have hosted the Eden community. It could have lived in a brief (up to 1 million years) period of hot (~100 °C) ocean after a major meteor impact, or possibly could have existed in a hydrothermal system (see Box 2). In the transient hot-ocean case, hydrothermal systems would have offered protective settings even after the ocean cooled.

An alternate version (the ‘hyperthermophile Noah’ hypothesis) is that the universal ancestor was not necessarily hyperthermophile, but diversified from an unknown Eden into an early population that included some hyperthermophiles near hydrothermal systems. During the earliest Archaeal, it would have been likely that a major meteorite impact capable of heating the ocean to 100 °C would have hit the earliest community^{61,62}. Only hyperthermophiles could survive an impact catastrophe — the ‘impact bottleneck’⁶³. Perhaps two by-then-distinct lineages of descendants survived the bottleneck, one to lead to the bacteria and the other to the archaea. The ‘Noah’ was the last common pre-impact ancestor of those organisms that survived, except possibly some viruses⁶⁴. Since then, the rainbow has shone over cool waters.

A third alternative is that the earliest evolution took place on Mars. There is no consensus whether life exists, or ever existed, on Mars, but early Mars, with then-vigorous volcanism, may have been a kinder, gentler habitat than Earth⁶². Large impacts would have occurred here too, but there was no deep ocean which could be vaporized to maintain a global steambath. Major impacts would have ejected many fragments from the surface of Mars, some of which would have landed on Earth. If

Mars did have life, possibly one or more cells survived impact on Mars, ejection, freezing in space, and transfer to and landing on Earth, where the cell line then survived later ocean-heating bottlenecks associated with impacts. Gene transfer is potential signal rather than noise as parts of the tree on the same planet would exchange with each other. A tenable ‘martian’ explanation of the origin of the Eucarya is that the ancestral eukaryote was a later, second martian transfer. In this hypothesis, the ancestors of the eukaryote stem cell continued to evolve on Mars, after the time the ancestor of the archaea and bacteria had been ejected to Earth. If a cell from this line were later ejected to Earth, the newly arrived distant cousin could have hosted a symbiotic union with the bacteria.

Both the notional ‘hyperthermophile Eden’ and ‘Noah’ last common ancestors discussed here must have been DNA-based and chemically sophisticated, possessing many of the basic housekeeping proteins. But these might not have been the original replicating organisms. The variety of suggestions about the setting of the origin of life is wide and unconstrained, beyond the scope of this review. Where was Eden? Possible birthplaces and first habitats range from ancestry in an aerosol, to a cold pond under a glacier, to a small, warm pond near a hydrothermal system⁶⁵, and many other alternatives.

The notion of a universal phylogenetic tree that crystallized into distinct branches has been criticized^{53,66} on the grounds that lateral gene transfer between contemporaneous organisms may have been on a much wider scale than is implied by a tree. There is clear evidence today for lateral transfers of genetic information. However, such

transfers between widely separated organisms would rarely be advantageous: for example, gaining information about photosynthesis would be of little use to a bacterium that lived deep in mud. Perhaps a better model is not a tree, even a mangrove, but a braided stream delta, with much cross-over, yet still clearly defined flows from the main distributaries. A closer analogy is the origin of languages: there are many parallels between the evolution of eukaryotes and the chimaera that amalgamated a Saxon root with French, Latin and Greek implants, and added doses of Arabic to Zulu, to make English.

Molecular phylogeny can be calibrated by reference to the geological record^{29,67}. To summarize, given the scale of the early bombardment, it is unlikely that Earth could have been a permanent habitation before about 4.2 or even 4.0 Gyr or less⁶¹. If the evidence from the Isua belt³⁴ does indeed record life, and the age is interpreted correctly, then life is perhaps up to 3.8 Gyr old, or more; moreover, this life possibly existed by anoxygenic photosynthesis, implying that considerable speciation had taken place by then. The last common ancestor would have existed long before that. By about 3.5 Gyr the Rubisco signature necessarily implies global oxygenic photosynthesis³² and the evolution of cyanobacteria. Ancestral eukaryotes appear to predate 2.7 Gyr (refs 48, 49).

Habitat occupation

It is possible that the earliest replicating system was an RNA molecule, capable of acting as a ribozyme and in some way self-assembling, which was ancestor of the ribosome. This RNA-world hypothesis⁶⁸ deserves special attention, especially as the 30S ribosomal subunit, which lies at the heart of the cell, is in effect a giant ribozyme supported by a scaffold of caring proteins^{69,70}. An RNA world may have lived in pores in rock around a hot spring where both vapour and liquid phases were present^{28,71}. Possibly the setting was subaerial, allowing both gas and fluid chemistry. The environment was presumably rich in phosphate, to allow a self-replicating molecule to gain access to essential phosphorus. Geologically, such environments tend to occur around alkaline volcanics, mainly but not all on continents. One possibility is an alkaline hot spring drawing fluid from an alkaline ultramafic body. Phosphates are associated with rocks such as carbonatites that are typically intruded in continental settings, but hydroxylapatite is known from rocks such as serpentinites⁷², which are similar to ocean-floor material and to altered komatiite, a lava common in the Archaean²⁸. Komatiite volcanoes would have built large Hawaii-like shields, perhaps hosting alkaline hydrothermal systems around subaerial volcanoes.

Many of the basic components of biochemical housekeeping may reflect ancestry in a hydrothermal system⁷³. These may include many of the metal proteins, especially those involving iron–sulphur, nickel, molybdenum, copper, cobalt and zinc, some of which, such as metal–nitrogen structures, may have condensed in alkaline settings. Heat-shock proteins possibly date from the time when the community before the last common ancestor lived on the fine line between starving (being too cold, and too far from the hot vent) and being cooked. Any property that could repair damage would be highly advantageous. The shaping properties of heat-shock proteins would preadapt them for use as chaperonins, helping protein folding.

The phylogenetic tree implies strongly that the first living community was not photosynthetic. It seems improbable that the sophisticated biochemistry needed for photosynthesis should spring out of nothing. Early Archaean life would have had access to redox contrast between a more oxidized atmosphere–ocean system, open at the top to space, and the more reduced fluids in contact with mantle-derived magmas. Sulphur offers the best opportunities. An early atmosphere rich in sulphur gases, with CO₂ partial pressure exceeding 2 bar, could have provided some warmth and protection against UV radiation⁷⁴. With water present in the air, and hence probably OH, volcanic sulphur dioxide would have been oxidized to sulphate⁷⁵, although this may have been slow on a glacial planet. Dissolved in water, sulphate would provide oxidation power for organisms to react against reduced species in hydrothermal fluids, such as hydrogen and methane, and in rock surfaces.

Perhaps the last common ancestor lived in a thin biofilm of cells near

Box 3

Gaia in the Archaean?

In the modern atmosphere, nitrogen is managed biologically by nitrifying and denitrifying bacteria, and has a lifetime of tens of millions of years. Carbon dioxide in air has a lifetime of centuries, whereas the lifetime of oxygen in the air is many millions of years, but they are obverse and reverse. The reduced carbon reservoir is on the surface (for example, in plants, peat and soil) and in sediment (in gas, oil and coal). Oxygen in the air is only a small store of the total oxidation power that has been created by the biological use of light¹⁰⁹. More oxidation power is stored in the long-term heritage of oxidized minerals in sediment, crust and mantle. Tectonic control on carbon storage can be important¹¹⁰. In one view, most of the modern atmosphere can be seen as a biological construct¹¹¹, although the basic geochemical controls should not be forgotten¹¹². Only argon is not managed by biological processes.

Although microbial and less productive than today, Archaean ecology¹¹³ used the same basic biochemical cycles as on the modern Earth. Purple bacteria and cyanobacteria did then what mitochondria and chloroplasts do today. Archaean methanogens did then what they do in bovine stomachs today. There may have been significant emission of dimethylsulphide by microbial plankton, important in the sulphur cycle then as today. Moreover, below the surface the sedimentary biosphere has always been microbial. From this comes the hypothesis that the Archaean atmosphere and greenhouse feedback loops were also biologically cycled^{28,111}. It is possible, however, that for periods the Archaean biosphere was 'upside-down', with the store of reductant in the air, and oxidant stored in the sediment (for example, as iron oxide)¹¹⁴. Prior to the oxic transition, hydrogen loss to space from methane photolysis could have been important in oxidant supply to the surface environment. After the transition the reduction in CH₄ and hence greenhouse warming could have caused global cooling¹¹⁵.

Hydrothermal water from liquid oceans cools plates, and the presence of this water in oceanic crust, when driven off during subduction into overlying mantle, causes melting and ultimately the formation of the granitoid rocks that, collectively, are the continents. No water, no continents¹⁶. If there is too little water so that the mid-oceanic ridge is exposed, oxygen is consumed in weathering. Too much water, and the flooded continents never get weathered. Liquid water, maintained by the regulation of the atmospheric greenhouse, has controlled the presence of continents and the functioning of plate tectonics. Has the day-to-day regulator of the greenhouse, and hence water, been life?

a hot vent, surviving on the redox contrast between slightly more oxidized water and slightly more reduced substrate, and on the difference between warm reduced water from vents and slightly more oxidized ambient water. Dead and dying cells would inevitably accumulate under the biofilm, creating a potential habitat-niche in recycling the reduced organic matter. Thus once a biofilm existed, niches would inevitably form and be filled by evolutionary divergence. Evolution works by tinkering with the available equipment⁷⁶, adapting existing organs to new purposes.

Archaean hydrothermal settings (Fig. 1) would have been varied^{42,73}. In deep water, mid-ocean ridge volcanism may have been much more active than today, with abundant vent fields including 'black smokers' and hydrothermal deposits rich in, for example, manganese, iron, copper, zinc and sulphur. Widespread volcanoes erupted komatiite, possibly forming shields similar to Hawaii today, but lower and much wider, hosting subaerial hot springs. If plate tectonics operated, subduction volcanism (comparable to that in Japan today) would also have occurred, hosting subaerial and subaqueous hydrothermal systems, with fluids containing, for example, copper, molybdenum and zinc.

Methanogens are deeply rooted on the standard tree, and it is possible to imagine an early biosphere inhabited by sulphate reducers that exploited the oxidation contrast between the air–water system and more reduced rock-derived fluids. These primary producers would be underlain by methanogenic recyclers^{77,78}. Such a biota would have produced surplus methane, which may, if the system was productive enough, have had global consequences by greenhouse warming. If the $\text{CH}_4\text{:CO}_2$ ratio were high enough, a biogenic, methane-rich smog layer^{23,24} might have formed that blocked UV light.

Origin of photosynthesis

There are various hypotheses for the evolution of photosynthesis^{79–83}. One plausible model is a sequence starting with accidental use of pigments by simple organisms living in a setting where local chemical disequilibrium is easily and accidentally exploited, leading to preadaptation that allowed cells to exploit light as an additional source of energy in anoxygenic photosynthesis, and finally the transition to full dependence on photosynthesis.

It was thought, on biochemical grounds, that chlorophyll biosynthesis predated bacteriochlorophyll, but recent evidence disputes this⁸³. The appearance of anoxygenic (bacteriochlorophyll) photosynthesis would have made shallow-level and subaerial hydrothermal systems much more productive. Anoxygenic photosynthesis exploits light in the longer visible and near-infrared spectrum. The specific light wavelengths used depend on the type of bacterium and setting in a mat^{79,84}. Purple bacteria have a wide spectrum of absorption, including bacteriochlorophylls absorbing at 900 or >1,000 nm, whereas green bacteria use bacteriochlorophylls that have absorption maxima around 750 nm in living cells. Anoxygenic photosynthesis uses a variety of electron donors in different bacteria, including hydrogen, hydrogen sulphide, sulphur and various organic chemicals. Some cyanobacteria can use sulphide in anoxygenic photosynthesis⁸⁵. Anoxygenic photosynthesis could have evolved in a bacterium using infrared thermotaxis⁸². This preadaptation, useful in a deep-water setting near hot hydrothermal vents, could have allowed a bacterium that drifted into shallow water to utilize sunlight and occupy mesothermophile habitats.

In contrast, oxygenic photosynthesis uses visible light in more energetic wavelengths, and has H_2O as the electron donor; Rubisco then helps capture carbon from CO_2 in the atmosphere–ocean system. The family of photosystem II reaction centres, using pigments and quinones as electron acceptors, is found in purple bacteria and in *Chloroflexus aurantiacus*, a green bacterium that may be from a line of great antiquity, as well as in cyanobacteria and chloroplasts. The photosystem I family of reaction centres, using iron–sulphur centres as electron acceptors, occurs in green sulphur bacteria, cyanobacteria and chloroplasts. The involvement of both photosystems in oxygenic photosynthesis indicates an origin from genetic transfer between cooperating or closely juxtaposed cells, each using anoxygenic photosynthesis.

A key component in oxygenic photosynthesis is the oxygen-evolving complex that is based on a manganese complex exploiting a transition from Mn_4O_4 to Mn_4O_6 . To the geologist, the involvement of manganese immediately suggests the vicinity of a vent of a hydrothermal system, but the environment needs to be oxygen-rich. The complex might have developed in a photosynthetic mat from manganese–catalase, perhaps to handle excess peroxide⁸¹, or as a toxic weapon to use against competitors, or both.

The evolution of the structure of microbial mats may have paralleled the evolution of photosynthesis, with newer forms progressively claiming occupation of the more productive but more dangerous uppermost level in the mats, where light was brighter⁸⁶. In this model, early pre-photosynthetic biofilms that were hyperthermophile and chemotrophic would have had bacterial sulphate-processors on top, underlain by archaea that recycled redox power. Anoxygenic photosynthetic mats would have added a top layer of bacterial photosynthesizers, introducing a new source of reduction power. This would have allowed occupation of the mesothermal outer perimeters of hydrothermal pools, and then the open environment away from

volcanism (Fig. 1). Finally in this model, before 3.5 Gyr, cyanobacteria brought a new occupant to the top layer. Their evolution, possibly as a bacterial chimaera formed from a genetic exchange between interdependent purple and green bacteria living on the redox boundary of a microbial mat, created an organism that could live freely on the planet, wherever water, light and CO_2 were present (Fig. 1). Coupled with their nitrogen-fixing ability, this would have allowed an enormous expansion of the biosphere. Whereas before the community away from hydrothermal systems was limited probably to sulphate-reducers depending on available sulphate, now life could spread widely (and the sulphate supply would rise too).

Is Rubisco a 'qwerty' enzyme?

Where CO_2 is in excess, as in the air, Rubisco⁸⁷ preferentially selects ^{12}C . For 3.5 Gyr, this isotopic signature in organic carbon, and the reciprocal signature in inorganic carbonate, has recorded Rubisco's role in oxygenic photosynthesis as the main link between atmospheric and biomass carbon³². But Rubisco itself may long predate oxygenic photosynthesis, as many non-photosynthetic microaerobic and aerobic bacteria use it. Unlike the many enzymes whose efficiency has been so honed by the aeons as to approach 100% (for example, catalase), Rubisco works either as carboxylase or oxygenase in photosynthesis and photorespiration⁸⁸. This apparent 'inefficiency', capable of undoing the work of the photosynthetic process, is paradoxical, yet fundamental to the function of the carbon cycle in the biosphere. Without it, the amount of CO_2 in the air would probably be much lower.

It is possible that Rubisco is not subject to evolutionary pressures because it has a monopoly. The qwerty keyboard, which is the main present link between humanity and the silicon chip, may be a parallel: legend is that qwerty was designed to slow typists' fingers so that the arms of early mechanical typewriters would not jam. It is among the worst, not the best, of layouts, and only minor evolution occurred (English has Y where German has Z). Perhaps the same applies to Rubisco: if so, genetic engineering to improve Rubisco might lead to a productivity runaway that removes all atmospheric CO_2 .

Origin of Eucarya

Late-Archaeon (2.7 Gyr) rocks contain molecules that suggest not only the presence of cyanobacteria but also eukaryotes^{48,49}. Eukaryotes may have evolved slowly, from a parental stem that symbiotically incorporated chloroplasts (which are descended from cyanobacteria) and then mitochondria (probably descended from α -proteobacteria). In this model the amitochondrial eukaryotes would be primitive. However, it is possible that the incorporation of mitochondria was synchronous with the origin of the eukaryote nucleus⁸⁹. In the 'hydrogen hypothesis'⁹⁰, a symbiotic partnership may have become a union between anaerobic hydrogen-dependent archaea and heterotrophic proteobacteria capable of producing molecular hydrogen through anaerobic fermentation. An alternative, but not necessarily exclusive, hypothesis⁹¹ is that an anaerobic archaeon could have evolved the ability to survive in more oxidizing environments (for example, near to oxygenic cyanobacteria) by incorporating symbiotic respiring proteobacteria. The origin of chloroplasts may have been a single ancestral cyanobacterium, but this is not proven⁹². Perhaps the eukaryote-creating event also incorporated mitochondria and chloroplasts simultaneously, in a single accident, possibly in a symbiotic consortium living on the redox boundary in a microbial mat.

The oxygen debate

The debate about the oxidation state of the Archaean atmosphere is vigorous, with strong proponents both of the 'oxic'⁹³ and not-oxic⁹⁴ viewpoints. Certain facts are available. In the early geological record, sulphate, although rare, does occur. Apparently evaporitic at first, it is present in ~3.5-Gyr rocks^{95,96}. Deposition of sulphate implies localized non-reducing conditions (although not necessarily the presence of free oxygen), at least at 3.5 Gyr. But many Archaean rocks contain apparently detrital pyrite, siderite and uraninite^{94,97}, minerals that are

difficult to transport far in oxidizing settings. Rocks widely recognized as redbeds do not exist before about 2.2 Gyr (there are some older exceptions, but these may have been oxidized later under Cretaceous or Tertiary erosion surfaces), and interpretation of palaeosols also suggests oxidizing conditions after this time.

Models of the chemistry of the early atmosphere, into which abundant CO₂ and sulphur gases would have been emitted by volcanism, suggest that oxygen was present, but at low partial pressures⁹⁸. Given the probable oxidation state of the mantle and thus the degassing of SO_x gases and subsequent presence of oxidant in the high atmosphere⁷⁵, as well as a possible supply of relict oxygen after hydrogen loss to space, it is likely that the supply of sulphur gases was adequate to support early chemotrophic life. Such life would have reacted chemical species from the relatively oxidized atmosphere–ocean system, bombarded by light and open to space, with more reduced mantle-exchanged hydrothermal fluids⁹⁸. The planet may have been covered mainly in ice, except for open-water leads⁹⁹, unless the CH₄ content of the air was high enough to sustain a methane greenhouse²⁴.

The appearance of oxygenic photosynthesis before 3.5 Gyr (ref. 32) provided a source of atmospheric oxidation power that would have increased the productivity of chemotrophic life. The nitrogen cycle may be of similar antiquity. Nitrogenase, which after Rubisco is the next most important enzyme (and is perhaps also a qwerty enzyme), consists of an iron protein and a molybdenum–iron protein that includes a 4Fe–4S cluster and a Mo–3Fe–3S cluster. The presence of molybdenum, iron and sulphur suggest a hydrothermal heritage, perhaps originally for dealing with ammonia in a reducing setting, as nitrogenase is inhibited rapidly by oxygen.

Life spreads instantly on a geological timescale, and immediately the modern carbon and nitrogen cycles were in place the atmosphere would have become biologically ruled by kinetics and disequilibrium, not sustainable equilibrium (see Box 3). There were huge reservoirs of reductant (for example, reduced iron, sulphide and organic debris in sediment), but the biosphere is inflationary in that it sequesters oxidant and reductant and exploits the possibilities of cycling between them. A thin layer of life can divide sharply contrasted redox reservoirs.

The abundance of Archaean ironstones implies that transport of iron took place from the source to the place of deposition. But only Fe²⁺ species are soluble, indicating that reducing conditions were required. Vast bacterial blooms could have produced ironstones, perhaps where deep, reduced water met shallow-level, oxidized water. The objection to this idea is that there is little organic carbon in ironstones: deposition may have been inorganic. However, biological iron transport could have been important in a microbial world. Bacteria could acquire iron in local reducing settings, such as near hydrothermal vents or in soil, and then pass the iron through the biomass by predation or by recycling dead bodies (but see Box 3 for Walker-world, the upside-down biosphere).

Global oxygen production has probably been of the same order of magnitude as today (to a factor of 10) for at least 3.5 Gyr, but the oxygen level in the atmosphere does not depend on production alone. Consider a bathtub (the atmosphere) with a running tap (oxygen production by photosynthesis). The level in the tub depends not so much on the flow from the tap but on the plug. If the plug is out, there will be little water in the tub even if the tap is full on; if it is in, the tub will eventually fill to the overflow limit, even if the tap only drips. The air may be similarly regulated. If the oxygen level increased sharply around 2.2 Gyr, possibly the appearance of complex eukaryotes may have been involved. The cellular cybernetic switch between mitochondria and chloroplasts¹⁰⁰ may control the link between photosynthesis, CO₂ and nitrogen fixation, in partnership with the ability of Rubisco to reverse its function^{87,88,101} as CO₂/O₂ ratios change. There may have been major excursions from this simple picture of a planet with a microaerobic early atmosphere that switched to oxic air after 2.2 Gyr. For example, at times global methane production by archaea may have been highly significant, with events when the atmospheric methane burden was high^{23,24,102}.

The evolution of the sulphur cycle remains controversial. Some isotopic evidence^{103,104} suggests that microbial fractionation of sulphur was limited in the Archaean, implying that sulphate concentrations were low. But other isotopic evidence⁵⁰ implies that the full sulphur cycle evolved earlier, which would be expected if the ‘standard’ molecular phylogeny⁵² is correct with respect to sulphur processing. It is possible that the microbial diversity of sulphur handlers was present early on, but only became widespread much later¹⁰⁵.

Afterview

On Earth, life probably dates from 3.8 Gyr or before, but life may have existed earlier on Mars or even on Venus or an outer moon and been translated to Earth by meteorite. We still have little idea how, when or where life began. The notion that life began in a hydrothermal setting is extremely attractive, but the evidence is circumstantial and can be compared with delving into such records as there are in Massachusetts of the Mayflower, to discern the origins of the English language.

The debate about life’s origins has deep resonance in our society. Those who work in this field frequently find their search challenged in assaults on empirical natural science. Judaeo–Christian thought must accept convincing evidence from nature; denial is both destructive of faith and dangerous to science. To find the fragments of fact, and to attempt to understand them, is a powerful response to the Creationist heresy. Not only fact and honest interpretation, but also orthodox theological argument reject Creationism: much Jewish and Christian thinking agrees with the anonymous writer of the epistle to the Hebrews, Peter and Augustine in the view that the Biblical Day is a wider concept than the 24-hour rotation of the Earth. The Seventh Day is lasting. The author of Job and Paul both challenge us to search nature, although we may not find the answer.

In summary, the best evidence is that life has been present on Earth since about 3.8 Gyr or earlier, and for at least the past 3.5 Gyr the main biochemical carbon cycle has been operating. But whether Earth is alone as a planet of life remains an open question. □

- Newman, J. H. *Apologia Pro Vita Sua* (Longmans, London, 1980).
- Wetherill, G. W. Formation of the Earth. *Annu. Rev. Earth Planet. Sci.* **18**, 205 (1990).
- Ahrens, T. J. in *Origin of the Earth* (eds Newsom, H. E. & Jones, J. H.) 211–227 (Oxford Univ. Press, New York, 1990).
- Lewis, J. S. & Prinn, R. G. *Planets and their Atmospheres* (Academic, Orlando, 1984).
- Hunt, D. M. Atmospheric evolution of the terrestrial planets. *Science* **259**, 915–920 (1993).
- Watson, A. J., Donahue, T. M. & Kuhn, W. R. Temperatures in a runaway greenhouse on the evolving Venus. *Earth Planet. Sci. Lett.* **68**, 1–6 (1984).
- Bell, D. R. & Ihinger, P. D. The isotopic composition of hydrogen in nominally anhydrous mantle minerals. *Geochim. Cosmochim. Acta* **64**, 2109–2118 (2000).
- Yung, Y., Wen, J.-S., Moses, J. I., Landry, B. M. & Allen, M. Hydrogen and deuterium loss from the terrestrial atmosphere: a quantitative assessment of non-thermal escape fluxes. *J. Geophys. Res.* **94**, 14971–14989 (1989).
- Trieloff, M., Kunz, J., Clague, D. A., Harrison, D. & Allegre, C. J. The nature of pristine noble gases in mantle plumes. *Science* **288**, 1036–1038 (2000).
- Carr, M. *Water on Mars* (Cambridge Univ. Press, Cambridge, 1996).
- Halliday, A. N. Terrestrial accretion rates and the origin of the Moon. *Earth Planet. Sci. Lett.* **176**, 17–30 (2000).
- Melosh, H. J. in *Origin of the Earth* (eds Newsom, H. E. & Jones, J. H.) 69–83 (Oxford Univ. Press, Oxford, 1990).
- Milton, J. *Nature Unimpaired by Time* (Latin verse, trans. W. Cowper, 1791) in *The Poetical Works of William Cowper* (Frederick Warne, London, New York, 1908).
- Wilde, S. A. *et al.* Evidence from detrital zircons for the existence of continental crust and oceans on the Earth 4.4 Gyr ago. *Nature* **409**, 175–178 (2001).
- Stern, R. A. & Bleeker, W. Age of the world’s oldest rocks refined using Canada’s SHRIMP: the Acasta Gneiss Complex, Northwest Territories, Canada. *Geosci. Can.* **25**, 1, 27–31 (1998).
- Campbell, I. H. & Taylor, S. R. No water, no granites, no oceans, no continents. *Geophys. Res. Lett.* **10**, 1061–1064 (1983).
- McKay, D. S. *et al.* Search for past life on Mars: possible relic biogenic activity in Martian meteorite ALH84001. *Science* **273**, 924–930 (1996).
- Kasting, J. F., Whitmire, D. P. & Reynolds, R. T. Habitable zones around main sequence stars. *Icarus* **101**, 108–128 (1993).
- Lewis, C. S. *Perelandra (Voyage to Venus)* (John Lane the Bodley Head, London, 1943).
- Zuber, M. T. *et al.* Internal structure and early thermal evolution of Mars from Mars global surveyor topography and gravity. *Science* **287**, 1788–1793 (2000).
- Sleep, N. H. & Zahnle, K. Refugia from asteroid impacts on early Mars and the early Earth. *J. Geophys. Res.* **103**, 28529–28544 (1998).
- Sagan, C. & Chyba, C. The early Sun paradox: organic shielding of ultraviolet-labile greenhouse gases. *Science* **276**, 1217–1221 (1997).
- Zahnle, K. Photochemistry of methane and the formation of hydrocyanic acid (HCN) in the Earth’s early atmosphere. *J. Geophys. Res.* **91**, 2819–2834 (1986).

24. Pavlov, A., Kasting, J. F., Brown, L. L., Rages, K. A. & Freedman, R. Greenhouse warming by CH₄ in the atmosphere of early Earth. *J. Geophys. Res.* **105**, 11981–11990 (2000).
25. Kress, M. E., Zahnle, K. & McKay, C. P. Impact production of CH₄ on early Earth and Mars. *EOS* **81**, F809 (2000).
26. Drake, M. J. Accretion and primary differentiation of the Earth. *Geochim. Cosmochim. Acta* **64**, 2363–2370 (2000).
27. Schidlowski, M. in *Instruments, Methods and Missions for Astrobiology* (Proc. Soc. Photo-Opt. Instrument. Engs **3441**) 149–157 (Bellingham, WA, 1998).
28. Nisbet, E. G. *The Young Earth: An Introduction to Archaean Geology* (Cambridge Univ. Press, Cambridge, 1987).
29. Nisbet, E. G. & Fowler, C. M. R. Some liked it hot. *Nature* **382**, 404–405 (1996).
30. Nutman, A. P., Mojzsis, S. J. & Friend, C. R. L. Recognition of >3850 Ma water-lain sediments and their significance for the early Earth. *Geochim. Cosmochim. Acta* **61**, 2475–2484 (1997).
31. Schidlowski, M. A 3,800 million-year old record of life from carbon in sedimentary rocks. *Nature* **333**, 313–318 (1988).
32. Schidlowski, M. & Aharon, P. in *Early Organic Evolution: Implications for Mineral and Energy Resources* (eds Schidlowski, M. *et al.*) 147–175 (Springer, Berlin, 1992).
33. Mojzsis, S. J. *et al.* Evidence for life on Earth 3800 million years ago. *Nature* **384**, 55–59 (1996).
34. Rosing, M. T. ¹³C-depleted carbon in >3700 Ma seafloor sedimentary rocks from West Greenland. *Science* **283**, 674–676 (1999).
35. Awramik, S. M. in *Early Organic Evolution: Implications for Mineral and Energy Resources* (eds Schidlowski, M. *et al.*) 435–439 (Springer, Berlin, 1992).
36. Lowe, D. R. Abiological origin of described stromatolites older than 3.2 Ga. *Geology* **22**, 387–390 (1994).
37. Buick, R., Dunlop, J. S. R. & Groves, D. I. Stromatolite recognition in ancient rocks: an appraisal of irregularly laminated structures in an early Archaean chert-barite unit from North Pole, Western Australia. *Alcheringa* **5**, 161–181 (1981).
38. Walsh, M. M. Microfossils from the early Archaean Onverwacht Group, Barberton Mountain land, South Africa. *Precambrian Res.* **54**, 271–293 (1992).
39. Schopf, J. W. & Packer, B. M. Early Archaean (3.3 billion to 3.5 billion year old) microfossils from Warrawoona Group, Australia. *Science* **237**, 70–73 (1987).
40. Westall, F. de Wit, M., Dann, J., van der Gaast, S. de Ronde, C. & Gerneke, D. Early Archaean fossil bacteria and biofilms in hydrothermally-influenced sediments from the Barberton greenhouse belt, South Africa. *Precambrian Res.* (in the press).
41. Rasmussen, R., Filamentous microfossils in a 3,235-million-year-old volcanogenic massive sulphide deposit. *Nature* **405**, 676–679 (2000).
42. Nisbet, E. G. The realms of Archaean life. *Nature* **405**, 625–626 (2000).
43. Wilks, M. E. & Nisbet, E. G. Archaean stromatolites from the Steep Rock Group, N. W. Ontario, Canada. *Can. J. Earth Sci.* **22**, 792–799 (1985).
44. Beukes, N. J. & Lowe, D. R. Environmental control on diverse stromatolite morphologies in the 3000 Myr Pongola Supergroup, South Africa. *Sedimentology* **36**, 383–397 (1989).
45. Walter, M. R. in *Earth's Earliest Biosphere* (ed. Schopf, J. W.) 187–213 (Princeton Univ. Press, Princeton, 1983).
46. Martin, A., Nisbet, E. G. & Bickle, M. J. Archaean stromatolites of the Belingwe Greenstone belt, Zimbabwe (Rhodesia). *Precambrian Res.* **13**, 337–362 (1980).
47. Buick, R. The antiquity of oxygenic photosynthesis: evidence from stromatolites in sulphate-deficient Archaean lakes. *Science* **255**, 74 (1992).
48. Brocks, J. J., Logan, G. A., Buick, R. & Summons, R. E. Archaean molecular fossils and the early rise of eukaryotes. *Science* **285**, 1033–1036 (1999).
49. Summons, R. E., Jahnke, L. L., Hope, J. M. & Logan, G. A. 2-Methylhopanoids as biomarkers for cyanobacterial oxygenic photosynthesis. *Nature* **400**, 554–557 (1999).
50. Grassineau, N. V. *et al.* Antiquity of the biological sulphur cycle: evidence from S and C isotopes in 2.7 Ga rocks of the Belingwe belt, Zimbabwe. *Proc. R. Soc. Lond. B* **268**, 113–119 (2001).
51. Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *J. Theor. Biol.* **8**, 357–366 (1965).
52. Woese, C. R. Bacterial evolution. *Microbiol. Rev.* **51**, 221–271 (1987).
53. Doolittle, W. F. Uprooting the tree of life. *Sci. Am.* **72**, 77 (February 2000).
54. Woese, C. R. The universal ancestor. *Proc. Natl Acad. Sci. USA* **95**, 6854–6859 (1998).
55. Graham, D. E., Overbeek, R., Olsen, G. J. & Woese, C. R. An archaeal genomic signature. *Proc. Natl Acad. Sci. USA* **97**, 3304–3308 (2000).
56. Pace, N. R. A molecular view of biodiversity and the biosphere. *Science* **276**, 734–740 (1997).
57. Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposals for the domains Archaea, Bacteria and Eucarya. *Proc. Natl Acad. Sci. USA* **87**, 4576–4579 (1990).
58. Margulis, L. in *Life: Origin and Evolution* (ed. Folsome, C. E.) 101–110 (Freeman, New York, 1979). [Reprint from *Sci. Am.* (August 1971).]
59. Stetter, K. O. in *Evolution of Hydrothermal Ecosystems on Earth (and Mars?)* (eds Bock, G. R. & Goode, J. A.) 1–18 (CIBA Foundation Symposium **202**) (Wiley, Chichester, 1996).
60. Galtier, N., Tourasse, N. & Gouy, M. A non-hyperthermophilic common ancestor to extant life forms. *Science* **283**, 220–221 (1999).
61. Sleep, N. H., Zahnle, K. J., Kasting, J. F. & Morowitz, H. J. Annihilation of ecosystems by large asteroid impacts on the early Earth. *Nature* **342**, 139–142 (1989).
62. Sleep, N. H., Zahnle, K. & Neuhoff, P. S. Initiation of clement surface conditions on the earliest Earth. *Proc. Natl Acad. Sci. USA* (in the press).
63. Gogarten-Boeckels, M., Hilario, E. & Gogarten, J. P. The effects of heavy meteorite bombardment on the early evolution—the emergence of the three domains of life. *Origins Life Evol. Biosphere* **25**, 251–264 (1992).
64. Forterre, P. in *Frontiers of Life* (eds Tran Than Van, J., Tran Than Van, K., Mounoulou, J. C., Schneider, J. & McKay, C.) 221–233 (Gif-sur-Yvette Editions Frontieres, 1992).
65. Darwin, C. Some unpublished letters (1871) ed. Sir Gavin de Beer. *Notes Rec. R. Soc. Lond.* **14**, 1 (1959).
66. Doolittle, W. F. At the core of the Archaean. *Proc. Natl Acad. Sci. USA* **93**, 8797–8799 (1996).
67. Knoll, A. A new molecular window on early life. *Science* **285**, 1025–1026 (1999).
68. Gilbert, W. The RNA world. *Nature* **319**, 618 (1986).
69. Carter, A. P. *et al.* Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibodies. *Nature* **407**, 340–348 (2000).
70. Williamson, J. R. Small subunit, big science. *Nature* **407**, 306–307 (2000).
71. Nisbet, E. G. RNA and hydrothermal systems. *Nature* **321**, 206 (1986).
72. Mitchell, L., Faust, G. T., Hendricks, S. B. & Reynolds, D. S. The mineralogy and genesis of hydroxylapatite. *Am. Mineral.* **28**, 356–371 (1943).
73. Nisbet, E. G. & Fowler, C. M. R. in *Tectonic, Magmatic, Hydrothermal and Biological Segmentation of Mid-ocean Ridges* (eds MacLeod, C. J., Tyler, P. A. & Walker, C. L.) Geol. Soc. Lond. Spec. Pub. **118**, 239–251 (1996).
74. Kasting, J. F., Zahnle, K. J., Pinto, J. P. & Young, A. T. Sulfur, ultraviolet radiation, and the early evolution of life. *Origins Life Evol. Biosphere* **19**, 95–108 (1989).
75. Bekki, S. Oxidation of volcanic SO₂: a sink for stratospheric OH and H₂O. *Geophys. Res. Lett.* **22**, 913–916 (1995).
76. Jacob, F. in *Evolution from Molecules to Men* (ed. Bendall, D. S.) 31–67 (Cambridge Univ. Press, Cambridge, 1983).
77. DeLong, E. F. Resolving a methane mystery. *Nature* **407**, 577–579 (2000).
78. Boetius, A. *et al.* A marine microbial consortium apparently mediating anaerobic oxidation of methane. *Nature* **407**, 623–626 (2000).
79. Pierson, B. K. in *Early Life on Earth* (ed. Bengtson, S.) 161–180 (Nobel Symposium **84**) (Columbia Univ. Press, New York, 1994).
80. Blankenship, R. E. Origin and early evolution of photosynthesis. *Photosyn. Res.* **33**, 91–111 (1992).
81. Blankenship, R. E. & Hartman, H. The origin and evolution of oxygenic photosynthesis. *Trends Biochem. Sci.* **23**, 94–97 (1998).
82. Nisbet, E. G., Cann, J. R. & van Dover, C. L. Origins of photosynthesis. *Nature* **373**, 479–480 (1995).
83. Xiong, Fischer, W. M., Inoue, K., Nakahara & Bauer, C. E. Molecular evidence for the early evolution of photosynthesis. *Science*, **289**, 1724–1730 (2000).
84. Pringault, O., Kuhl, M., de Wit, R. & Caumette, P. Growth of green sulphur bacteria in experimental benthic oxygen, sulphide, pH and light gradients. *Microbiology* **144**, 1051–1061 (1998).
85. Cohen, Y., Jorgensen, B. B., Padan, E. & Shilo, M. Sulphide-dependent anoxygenic photosynthesis in the cyanobacterium *Oscillatoria limnetica*. *Nature* **257**, 489–492 (1975).
86. Nisbet, E. G. & Fowler, C. M. R. Archaean metabolic evolution of microbial mats. *Proc. R. Soc. Lond. B* **266**, 2375–2382 (1999).
87. Lorimer, G. H. The carboxylation and oxygenation of ribulose 1,5-bisphosphate: the primary events in photosynthesis and photorespiration. *Annu. Rev. Plant Physiol.* **32**, 349–383 (1981).
88. Lorimer, G. H. & Andrews, T. J. Plant photorespiration—an inevitable consequence of the existence of atmospheric oxygen. *Nature* **243**, 359 (1973).
89. Gray, M. W., Burger, G. & Lang, B. F. Mitochondrial evolution. *Science* **283**, 1476–1481 (1999).
90. Martin, W. & Muller, M. The hydrogen hypothesis for the first eukaryote. *Nature* **392**, 37–41 (1998).
91. Vellai, T. & Vida, G. The origin of eukaryotes: the difference between eukaryotic and prokaryotic cells. *Proc. R. Soc. Lond. B* **266**, 1571–1577 (1999).
92. Stiller, J. W. & Hall, B. D. The origin of red algae: implications for plastid evolution. *Proc. Natl Acad. Sci. USA* **94**, 4520–4525 (1997).
93. Ohmoto, H. When did the Earth's atmosphere become oxic? *Geochem. News* **93**, 12–13 (1997).
94. Holland, H. D. When did the Earth's atmosphere become oxic? A reply. *Geochem. News* **100**, 20–22 (1999).
95. Buick, R. & Dunlop, J. S. R. Evaporitic sediments of early Archaean age from the Warrawoona Group, North Pole, Western Australia. *Sedimentology* **37**, 247–277 (1990).
96. Barley, M. E. Volcanic, sedimentary and tectonostratigraphic environments of the ~3.46 Ga Warrawoona megasequence: a review. *Precambrian Res.* **60**, 47–67 (1993).
97. Rasmussen, R., & Buick, R. Oily old ores, evidence for hydrothermal petroleum generation in an Archaean volcanogenic massive sulphide deposit. *Geology* **27**, 115–118 (2000).
98. Kasting, J. F. Earth's early atmosphere. *Science* **259**, 920–925 (1993).
99. Sleep, N. H. & Zahnle, K. Carbon dioxide cycling and implications for climate on ancient Earth. *J. Geophys. Res.* **106** (in the press).
100. Joshi, H. M. & Tabita, F. R. A global two-way component signal transduction system that integrates the control of photosynthesis, carbon dioxide assimilation and nitrogen fixation. *Proc. Natl Acad. Sci. USA* **93**, 14515–14520 (1996).
101. Tolbert, N. E. in *Regulation of Atmospheric CO₂ and O₂ by Photosynthetic Carbon Metabolism* (eds Tolbert, N. E. & Preiss, J.) 8–33 (Oxford Univ. Press, Oxford, 1994).
102. Hayes, J. M. in *Early Life on Earth* (ed. Bengtson, S.) 220 (Nobel Symposium **84**) (Columbia Univ. Press, New York, 1994).
103. Habicht, K. S. & Canfield, D. E. Sulphur isotope fractionation in modern microbial mats and the evolution of the sulphur cycle. *Nature* **382**, 342–343 (1996).
104. Canfield, D. E. & Teske, A. Late Proterozoic rise in atmospheric oxygen concentration inferred from phylogenetic and sulphur-isotope studies. *Nature* **382**, 127–132 (1996).
105. Canfield, D. E., Habicht, K. S. & Thamdrup, B. The Archaean sulfur cycle and the early history of atmospheric oxygen. *Science* **288**, 658–661 (2000).
106. Wolery, T. J. & Sleep, N. H. in *Chemical Cycles in the Evolution of the Earth* (eds Gregor, C. B., Garrels, R. M., MacKenzie, F. T. & Maynard, J. B.) Vol. 3, 76–103 (Wiley, New York, 1988).
107. Cas, R. A. F. Submarine volcanism: eruption styles, products, and relevance to understanding the host-rock successions to volcanic-hosted massive sulphide deposits. *Econ. Geol.* **87**, 511–541 (1992).
108. Smith, J. V., Arnold, F. P., Parsons, I. & Lee, M. R. Biochemical evolution III: polymerisation on organophilic silica-rich surfaces, crystal-chemical modelling, formation of first cells, and geological clues. *Proc. Natl Acad. Sci. USA* **96**, 3479–3485 (1999).
109. Schidlowski, M. Early atmospheric oxygen levels: constraints from Archaean photoautotrophy. *J. Geol. Soc. Lond.* **141**, 243–250 (1984).
110. Des Marais, D. J. Tectonic control of the crustal organic carbon reservoir during the Precambrian. *Chem. Geol.* **114**, 303–314 (1994).
111. Lovelock, J. E. *Ages of Gaia* (Norton, London, 1988).
112. Walker, J. C. G. & Drever, J. I. in *Chemical Cycles in the Evolution of the Earth* (eds Gregor, C. B., Garrels, R. M., MacKenzie, F. T. & Maynard, J. B.) Vol. 2, 55–75 (Wiley, New York, 1988).
113. Nisbet, E. G. in *Early Precambrian Processes* (eds Coward, M. P. & Ries, A. C.) Geol. Soc. Lond. Spec. Pub. **95**, 27–51 (1995).
114. Walker, J. C. G. Was the Archaean biosphere upside down? *Nature* **329**, 710–712 (1987).
115. Catling, D. C., McKay, C. P. & Zahnle, K. J. The role of biogenic methane in the oxidation state of early Earth. *EOS* **81**, F809 (2000).

Acknowledgements

We thank many colleagues for comment (and some helpful disagreement), including W. Bleeker, J. Bowyer, R. Buick, N. Butterfield, D. Catling, F. Dyson, M. Fowler, N. Grassineau, B. Pierson, M. Schidlowski, C. Tickell and K. Zahnle. The work was supported by the Leverhulme Trust and NERC. EGN's contribution derives from a Macgregor Memorial Lecture sponsored by the Geological Society of Zimbabwe.

Life in extreme environments

Lynn J. Rothschild & Rocco L. Mancinelli

NASA Ames Research Center, Moffett Field, California 94035-1000, USA (e-mail: Lrothschild@mail.arc.nasa.gov; rmancinelli@mail.arc.nasa.gov)

Each recent report of liquid water existing elsewhere in the Solar System has reverberated through the international press and excited the imagination of humankind. Why? Because in the past few decades we have come to realize that where there is liquid water on Earth, virtually no matter what the physical conditions, there is life. What we previously thought of as insurmountable physical and chemical barriers to life, we now see as yet another niche harbouring 'extremophiles'. This realization, coupled with new data on the survival of microbes in the space environment and modelling of the potential for transfer of life between celestial bodies, suggests that life could be more common than previously thought. Here we examine critically what it means to be an extremophile, and the implications of this for evolution, biotechnology and especially the search for life in the Universe.

Normal is passé; extreme is chic. While Aristotle cautioned "everything in moderation", the Romans, known for their excesses, coined the word 'extremus', the superlative of *exter* ('being on the outside'). By the fifteenth century 'extreme' had arrived, via Middle French, to English. At the dawning of the twenty-first century we know that the Solar System, and even Earth, contain environmental extremes unimaginable to the 'ancients' of the nineteenth century. Equally marvellous is the detection of organisms that thrive in extreme environments. Macelroy¹ named these lovers ('philos' to the Greeks) of extreme environments 'extremophiles'.

The discovery of extreme environments and the organisms that inhabit them has made more plausible the search for life outside the Earth, and even the possibility of panspermia (the transport of life from one planet to another). The discovery of extremophiles has also put vitality into the biotech industry and dreams of stock options in the minds of field biologists. The discipline has exploded during the past decade, with several reviews published on extremophiles²⁻⁴, an increasing number of meetings held⁵, genomes sequenced and patents filed, and the launch of concerted funding programmes such as the US National Science Foundation and NASA's programmes in Life in Extreme Environments, Exobiology and Astrobiology, and the European Union's Biotechnology of Extremophiles and Extremophiles as Cell Factories⁶. Here we examine what it means to be an extremophile starting from first principles. As a result, we highlight extremophiles that are often overlooked, possibly because they are eukaryotes. We then focus on the significance of extremophile research to the search for life in the Universe, and conclude with a discussion of the future of extremophile research including their economic potential.

What is an extremophile?

An organism that thrives in an extreme environment is an extremophile; in more than one extreme it is a polyextremophile. Examples of the latter would include *Sulfolobus acidocaldarius*, an archaea that flourishes at pH 3 and 80 °C (Fig. 1). 'Extremes' include physical extremes (for example, temperature, radiation or pressure) and geochemical extremes (for example, desiccation, salinity, pH, oxygen species or redox potential) (Table 1). It could be argued that extremophiles should include organisms

thriving in biological extremes (for example, nutritional extremes, and extremes of population density, parasites, prey, and so on).

'Extremophile' conjures up images of prokaryotes, yet the taxonomic range spans all three domains. Although all hyperthermophiles are members of the Archaea and Bacteria, eukaryotes are common among the psychrophiles, acidophiles, alkaliphiles, piezophiles, xerophiles and halophiles (which respectively thrive at low temperatures, low pH, high pH, and under extremes of pressure, desiccation and salinity; see <http://www.astrobiology.com/extreme.html> for an overview). Extremophiles include multicellular organisms, and psychrophiles include vertebrates.

Although these characterizations seem straightforward, three philosophical issues need further exploration. First, what is 'extreme'? Perhaps 'extreme' is in the eyes of the beholder. It is clear that to a thermophile that dies at 21 °C and a piezophile that finds atmospheric pressure 'extreme', what determines an extremophily is based on definitions that are perhaps anthropocentric. There are two possibilities that are more scientifically tenable. The first is based on an evolutionary perspective — that is, the earliest



Figure 1 Congress Pool, Norris Geyser Basin, Yellowstone National Park, USA, where Tom Brock originally isolated *Sulfolobus acidocaldarius*. The average pH is 3 and the average temperature is 80 °C. Photo taken on 20 September 2000.

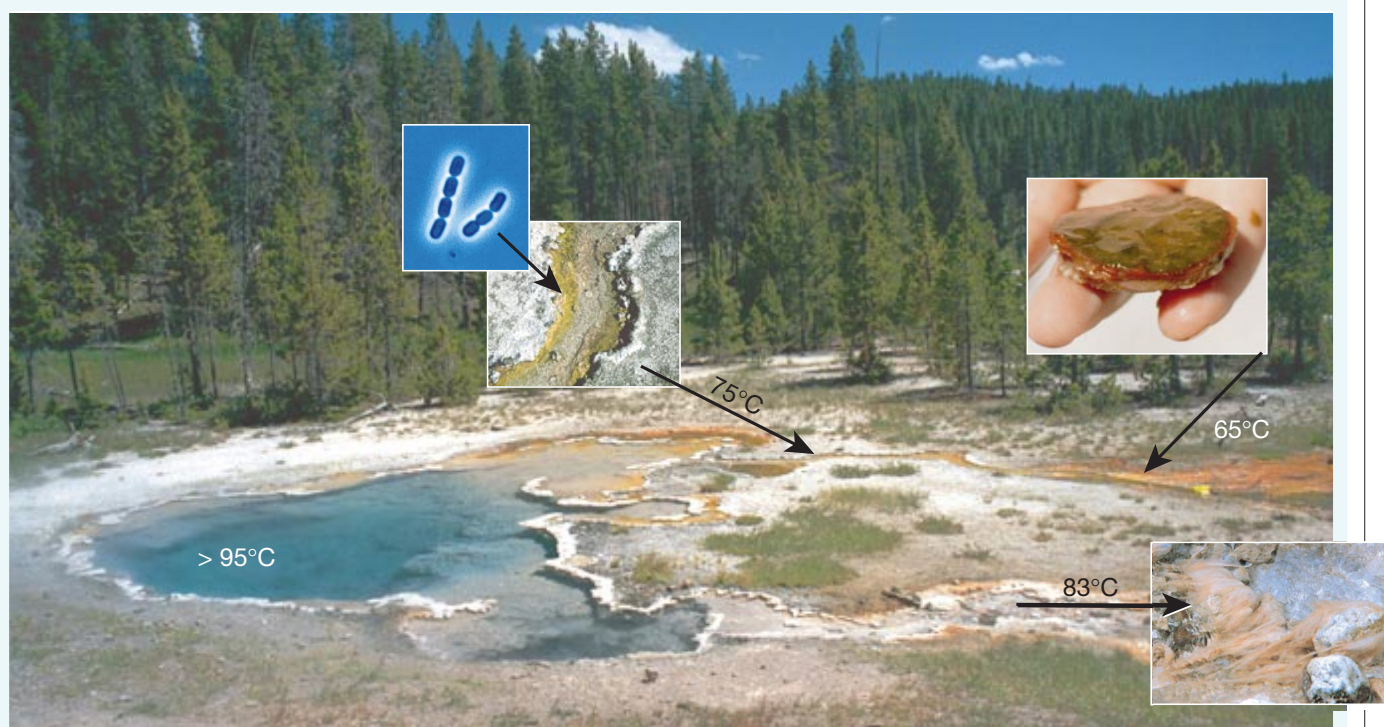


Figure 2 Octopus Spring, an alkaline (pH 8.8–8.3) hot spring in Yellowstone National Park, USA, is situated several miles north of Old Faithful geyser. The water flows from the source at 95 °C to an outflow channel, where it cools to a low of 83 °C. About every 4–5 minutes a pulse of water surges from the source raising the temperature as high as 88 °C. In this environment the pink filamentous *Thermocrinis ruber* thrives (lower right). Where the water cools to ≤ 75 °C, growth of photosynthetic organisms is permitted. The inset on left shows the growth of a thermophilic cyanobacterium, *Synechococcus*, tracking the thermal gradient across the channel. At 65 °C a more complex microbial mat forms with *Synechococcus* on the top overlaying other bacteria, including species of the photosynthetic bacterium *Chloroflexus* (upper right). The yellow object at 65 °C was part of an experimental set-up. Photo taken on 4 July 1999.

environment for life defines what is ‘normal’. If life arose in a high-temperature, anoxic hydrothermal vent, any environment that deviates from that is ‘extreme’. The second, which we favour, is based on a more objective, physical definition of ‘extreme’. This definition is congruent with the colloquial definition, with exceptions. All physical factors are on a continuum, and extremes in the conditions that make it difficult for organisms to function are ‘extreme’. For example, to maintain chemistry in an aqueous environment, cells need certain temperatures, pH and solutes, precise control over biomolecules and electric currents, and the ability to repair damage. There are certain conditions that will destroy biomolecules, such as desiccation, radiation and oxygen. Regarding the last of these conditions, oxygen forms reactive oxygen species that cause oxidative damage to nucleic acids, proteins and lipids^{7,8}. Thus, we and all other aerobes are extremophiles.

The second philosophical issue is ecological. Must an extremophile actually ‘love’ (remember ‘philos’) an extreme environment or can it merely tolerate it? In a practical sense the latter is clearly easier to determine experimentally, whereas in a biological sense the former has a certain biological and linguistic simplicity. In the last few decades of the twentieth century, numerous true extreme-loving organisms were found, thus permitting linguistic purity. But as a caveat, note that it is common for some environmental extremes (for example, radiation, vacuum or metal concentrations) to include organisms that tolerate rather than love the environment.

Third, does an organism have to be an extremophile during all life stages, and under all conditions? The bacterium *Deinococcus radiodurans*, the present gold-medallist of radiation resistance, is widely considered an extremophile par excellence. Yet, radiation resistance in *D. radiodurans* is severely diminished in stationary compared with logarithmic phase growth⁹, under increased concentrations of Mn^{2+} (ref. 10), with freezing or desiccation, and under nutrient-limited

conditions¹¹. Spores (for example, *Bacillus subtilis*), seeds and egg stages (for example, shrimp) are all far more resistant to environmental extremes than the vegetative forms. Trees, frogs, insects and fish can tolerate remarkably low temperatures during the winter as a result of seasonal shifts in physiology. Tardigrades (‘water bears’) in the tun state, can survive temperatures from -253 °C to 151 °C, X-rays, vacuum and, when in perfluorocarbon, pressures up to 600 MPa, almost 6,000 times atmospheric pressure at sea level¹².

Environmental extremes

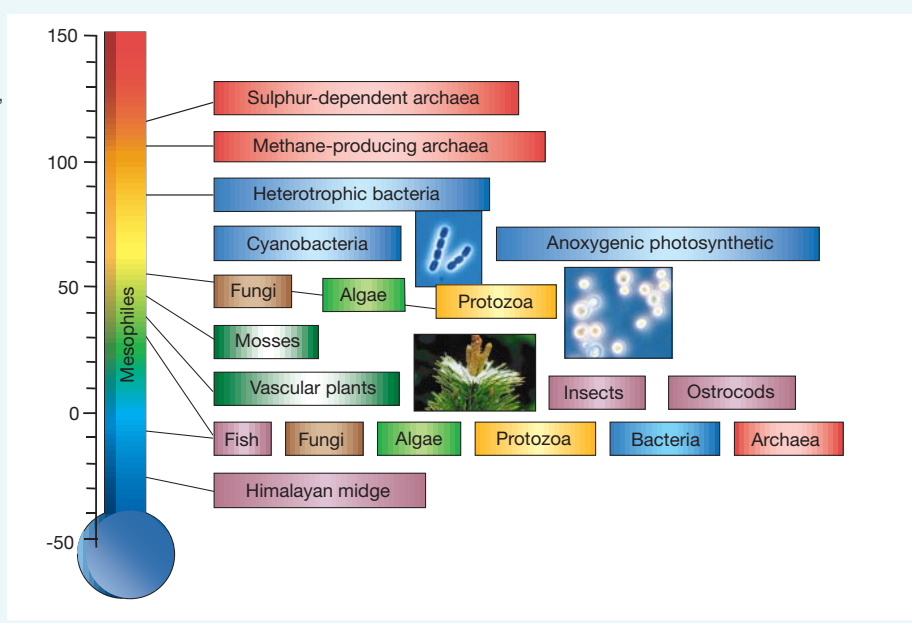
Liquid water is the *sine qua non* of life on Earth, and arguably any life in our Solar System¹³. Furthermore, life requires an input of energy, but must also be able to control energy flow. Redox chemistry is universal. As life is based on organic chemistry, such chemistry must be allowed to operate. An extremophile must either live within these parameters, or guard against the outside world in order to maintain these conditions intracellularly. With these rules in mind, we examine selected environmental parameters, summarized in Table 1.

Temperature

Temperature creates a series of challenges, from the structural devastation wrought by ice crystals at one extreme, to the denaturation of biomolecules at the other. The solubility of gases in water is correlated with temperature, creating problems at high temperature for aquatic organisms requiring O_2 or CO_2 . Temperatures approaching 100 °C normally denature proteins and nucleic acids, and increase the fluidity of membranes to lethal levels. Chlorophyll degrades above 75 °C, excluding photosynthesis (Fig. 2).

Yet, in nature thermal preferences range from hyperthermophilic¹⁴ (maximum growth >80 °C) to psychrophilic (maximum growth <15 °C). The most hyperthermophilic organisms are archaea, with *Pyrolobus fumarii* (Crenarchaeota), a nitrate-

Figure 3 Temperature limits for life. The highest and lowest temperature for each major taxon is given. Archaea are in red, bacteria in blue, algae in light green, fungi in brown, protozoa in yellow, plants in dark green and animals in purple.



reducing chemolithoautotroph, capable of growing at the highest temperatures of up to 113 °C (ref. 15). Hyperthermophile enzymes can have an even higher temperature optimum; for example, activity up to 142 °C for amylopullulanase¹⁶. There are thermophiles among the phototrophic bacteria (cyanobacteria, purple and green bacteria), eubacteria (*Bacillus*, *Clostridium*, *Thiobacillus*, *Desulfo-tomaculum*, *Thermus*, lactic acid bacteria, actinomycetes, spirochetes and numerous other genera) and the archaea (*Pyrococcus*, *Thermococcus*, *Thermoplasma*, *Sulfolobus* and the methanogens). In contrast, the upper limit for eukaryotes is ~60 °C, a temperature suitable for some protozoa, algae and fungi. The maximum temperature for mosses is lower by another 10 °C, for vascular plants it is about 48 °C, and for fish it is 40 °C, possibly owing to the low solubility of oxygen at high temperatures (Fig. 3).

Representatives of all major taxa inhabit temperatures just below 0 °C. Many microbes and cell lines can be preserved successfully at –196 °C (liquid nitrogen), but the lowest recorded temperature for active microbial communities is substantially higher, at –18 °C (ref. 17). Among animals, the Himalayan midge is active at –18 °C (ref. 18). Liquid water not only is a solvent for life as we know it, but also is important either as a reactant or product in most metabolic processes¹⁹. At low temperatures with nucleation, water freezes. The resulting ice crystals can rip cell membranes, and solution chemistry stops in the absence of liquid water. Freezing of intracellular water is almost invariably lethal. The only exception to this rule reported so far, outside of cryopreservation, is the nematode *Panagrolaimus davidi*, which can withstand freezing of all body water²⁰.

Radiation

Radiation is energy in transit, either as particles (for example, neutrons, electrons, protons, alpha particles or heavy ions) or electromagnetic waves (for example, gamma rays, X-rays, ultraviolet (UV) radiation, visible light, infrared, microwaves or radiowaves). Exceptional levels of radiation — sufficient to qualify for ‘extremophile’ status — rarely occur on the Earth naturally, but intense levels of UV and ionizing radiation are well-studied because of their importance to medicine, energy production, warfare and space travel. The dangers of UV and ionizing radiation range from decreased motility to inhibition of photosynthesis, but the most serious is damage to nucleic acids. Direct damage to DNA or indirect damage through the production of reactive oxygen species creates modified bases and single- and double-strand breaks.

The bacterium *D. radiodurans* is famous for its ability to withstand ionizing radiation (up to 20 kGy of gamma radiation) and UV

radiation (doses up to 1,000 J m⁻²), but this extraordinary resistance is thought to be a by-product of resistance to extreme desiccation²¹. Other organisms that can stand high levels of radiation are two *Rubrobacter* species²² and the green alga *Dunaliella bardawil*²³.

Pressure

Hominids evolved at an atmospheric pressure of 101 kPa (= 1 atmosphere = 1.013 bar), although our aquatic ancestors originated under hydrostatic pressure. Hydrostatic pressure increases at a rate of 10.5 kPa per metre depth, compared with 22.6 kPa per metre for lithostatic pressure. Pressure decreases with altitude, so that by 10 km above sea level, atmospheric pressure is almost a quarter of that at sea level. The boiling point of water increases with pressure, so water at the bottom of the ocean remains liquid at 400 °C. Because liquid water normally does not occur above ~100 °C, increased pressure can increase the optimal temperature for microbial growth, but usually by only a few degrees²⁴.

Pressure challenges life because it forces volume changes. Pressure compresses packing of lipids resulting in decreased membrane fluidity²⁵. If a chemical reaction results in an increase in volume, as most do, it will be inhibited by an increase in pressure²⁶. Although many organisms have adapted to very high pressures, a sudden change can be lethal, an effect only too well known to divers.

The Mariana trench (11° 22' N, 142° 25' E) is the world's deepest sea floor at 10,898 m, yet it harbours organisms that can grow at standard temperature and pressure. It also has yielded obligatory piezophilic species²⁷ that can grow at 70 to 80 MPa, but not below 50 MPa.

One component of pressure is gravity. Until now, organisms on Earth have, except for brief moments, lived at 1g. Space exploration will include extended periods in locations with gravity regimes different from our own: for example, launch vehicles (variable g), the International Space Station (microgravity), the Moon (0.17g) and Mars (0.38g). Although most of the concern with the effect of gravity have focused on human health, gravitational effects also have been found for microbes and include changes in biomass production, an increase in conjugation and changes in membrane permeability in *Escherichia coli*²⁸.

Desiccation

Water possesses many properties that seem to make it the essential solvent for life. It has high melting and boiling points with a wide temperature range over which it remains liquid, and a high dielectric

constant important for its solvent action. Water expands near its freezing point, and it forms hydrogen bonds. No other compound possesses all of these traits. Thus, water limitation is an extreme environment. Organisms that can tolerate extreme desiccation enter anhydrobiosis, a state characterized by little intracellular water and no metabolic activity. A variety of organisms can become anhydrobiotic, including bacteria, yeast, fungi, plants, insects, tardigrades, mycophagous nematodes and the shrimp *Artemia salina*^{29–32}.

Mechanisms of death due to anhydrobiosis include irreversible phase changes to lipids, proteins and nucleic acids such as denaturation and structural breakage through Maillard reactions, and accumulation of reactive oxygen species during drying, especially under solar radiation^{33–35}.

Salinity

Organisms live within a range of salinities, from essentially distilled water to saturated salt solutions. Osmophily refers to the osmotic aspects of life at high salt concentrations, especially turgor pressure, cellular dehydration and desiccation. Halophily refers to the ionic requirements for life at high salt concentrations. Although these phenomena are physiologically distinct, they are environmentally linked. Thus, a halophile must cope with osmotic stress. Halophiles include a range of microbes, but some archaea, cyanobacteria and the green alga *Dunaliella salina* can withstand periods in saturated NaCl.

pH

pH is defined as $-\log_{10}[\text{H}^+]$. Biological processes tend to occur towards the middle range of the pH spectrum, and intracellular and environmental pH often fall in this range (for example, the pH of sea water is ~8.2). However, in principle, pH can be high, such as in soda lakes or drying ponds, or as low as 0 ($[\text{H}^+] = 1 \text{ M}$) and below. Proteins denature at exceptionally low pH, which is what happens during the preparation of cerviche, the Latin American seafood dish ‘cooked’ in lime juice.

Acidophiles thrive at low pH (Fig. 4). Fish and cyanobacteria have not been found below pH 4, plants and insects below pH 2–3. Several unicellular eukaryotes do live below pH 1. The best characterized is the red alga *Cyanidium caldarium*³⁶, which has been described from nature at pH as low as 0.5, although its growth optimum in culture is pH 2–3 (ref. 37; Fig. 5). The green alga *Dunaliella acidophila* can also survive pH 0, with a sharp growth maximum at pH 1 (ref. 38). Three fungi, *Acontium cylatium*, *Cephalosporium* sp. and *Trichosporon cerebriae*, grow near pH 0 (ref. 39). Archaea have also been found flourishing under extreme acidity. The aerobic heterotrophs *Picrophilus oshimae* and *Picrophilus torridus* were isolated from Japanese soils permeated with solfataric gases, and had optimal growth at pH 0.7 and 60 °C (ref. 40). *Ferroplasma acidarmanus* has been described growing at pH 0 in acid mine drainage in Iron Mountain in California⁴¹, thriving in a brew of sulphuric acid and high levels of copper, arsenic, cadmium and zinc with only a cell membrane and no cell wall.

Alkaliphiles prefer high pH, which is an equally challenging environment. As with low pH, there is often a difference of 2 or more pH units between the internal and external milieu of the cell. Protons are

scarce, creating energetic hurdles for aerobic prokaryotes with a membrane-bound ATP synthase⁴². Representatives of all domains and kingdoms of eukaryotes are able to tolerate pH as high as ~11 (Fig. 4; refs 43, 44).

Oxygen

The Earth has been anaerobic throughout most of the history of life. Today organisms inhabit environments ranging from strictly anaerobic to aerobic. Aerobic metabolism is far more efficient than anaerobic, but the exploitation of oxygen metabolism has its costs. Oxidative damage resulting from the reduced forms of molecular oxygen, especially the hydroxyl radical, is extremely serious. Oxidative damage has been implicated in an array of health problems from ageing⁴⁵ to cancer⁴⁶, and has a range of consequences in nature (L.J.R., C.L. Wilson, N. Chough and R. I. Donaldson, unpublished results).

Reactive oxygen species are a pervasive threat. There is photochemical production of such species as H_2O_2 by UVA radiation (320–400 nm) within cells⁷, and metabolic production during aerobic metabolism and photosynthesis. Other endogenous sources of reactive oxygen species in eukaryotes include mitochondrial respiration (a significant source of O_2^-), cytochrome P450 metabolism of hydroperoxides (an important source of $^1\text{O}_2$ (singlet oxygen)), production of uric acid, and oxidative bursts used in fighting pathogens in animals and plants. Exogenous sources include the photochemical production of H_2O_2 in aquatic systems⁴⁷, and the production of the hydroxyl radical by ionizing radiation. The presence of oxygen can enhance radiation-induced DNA damage⁷.

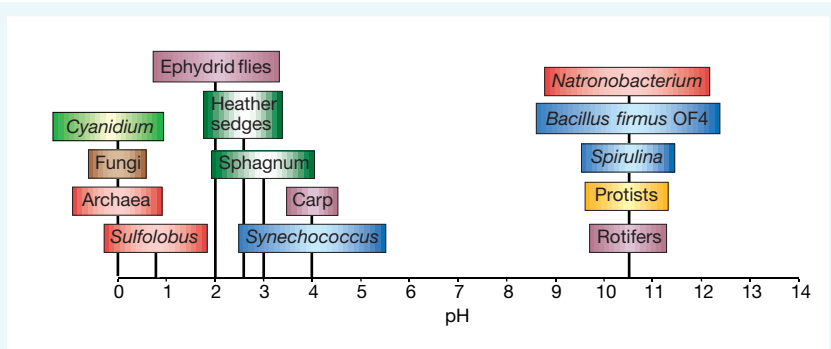
Other extreme conditions

A little creative thinking suggests other physical and chemical extremes not considered here. These include extremes in gas composition (*Cyanidium* grow in media ventilated with pure CO_2 (ref. 48)), redox potential, toxic or xenobiotic (synthetic) compounds, and heavy metal concentration⁴⁹. There are organisms that can live immersed in high levels of organic solvents⁵⁰. The electric eel (*Electrophorus electricus*) can produce, and thus must tolerate, strong electric currents.

How do they do it?

It is critical for an organism to maintain function, and the easiest approach to achieve this is to keep the external environment out. For example, *Cyanidium caldarium* and *Dunaliella acidophila* are found at pH 0.5, yet have near neutral cytoplasm^{38,51}, although this implies that extracellular proteins are acid-tolerant. The next step is to remove the problem as fast as possible. Heavy metal-resistant bacteria use an efflux pump to remove, for example, zinc, copper and cobalt, but not mercury, which is volatilized⁴⁹. If it is impossible to keep the environment out, evolutionary responses entail protective mechanisms, altering physiology or enhancing repair capabilities. Research has focused so far on three key classes of biomolecules: nucleic acids, membrane lipids and proteins. For nucleic acids, function and structure are linked inextricably. DNA is especially vulnerable to high temperature, radiation, oxidative damage and

Figure 4 pH limits for life. Examples of known pH limits for life are shown. Archaea are in red, bacteria in blue, algae in light green, assorted protists in yellow, fungi in brown, plants in dark green and animals in purple.



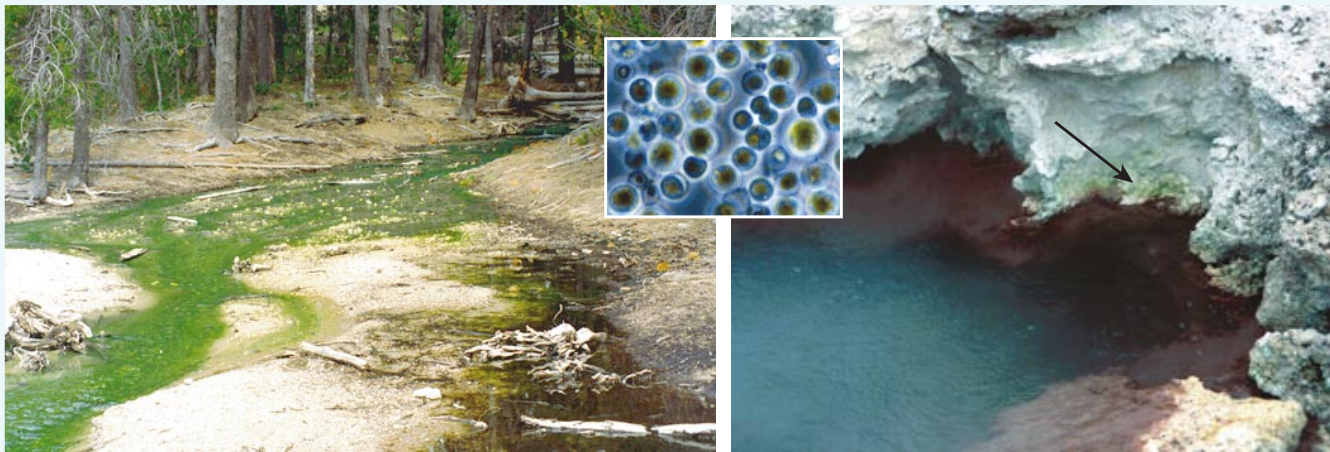


Figure 5 *Cyanidium caldarium*, Norris Geyser Basin, Yellowstone National Park, USA. The red alga *C. caldarium* can grow in the laboratory at a range of pH and temperature, but seems to be a superior competitor in nature at pH 3.3–3.5 and ~42 °C. On the left is Nymph Creek and on the right is Iron Spring. When the steam from Iron Spring cools to ~50 °C, *Cyanidium* can colonize the moist rock.

desiccation. This can lead either to convergence or to multiple ways to solve the problem of living in a particular environment. Understanding the alternatives by using extremophiles on Earth as a sample should help us understand evolutionary processes on Earth, predict them elsewhere, and be useful in commercial exploitation of extremophiles.

High-temperature adaptation

High temperature increases the fluidity of membranes. To maintain optimal membrane fluidity the cell must adjust the composition of the membrane including the amount and type (for example, saturated versus unsaturated) of lipids. Temperature also effects the structure and function of proteins⁵². Ways that proteins have evolved to cope with high temperatures include increasing ion-pair content, forming higher-order oligomers and decreasing flexibility at room temperature. Decreasing the length of surface loops is also known, in particular those loops that connect elements of secondary structure, optimize electrostatic and hydrophobic interactions, and exchange amino acids to increase internal hydrophobicity and helix propensity of residues in α -helices.

DNA at high (>70 °C) temperatures is subject normally to denaturation and chemical modification, yet the DNA of hyperthermophiles such as *Pyrococcus furiosus* is known to be more stable *in vivo* than that of a mesophile such as *Escherichia coli*⁵³. Monovalent and divalent salts enhance the stability of nucleic acids because these salts screen the negative charges of the phosphate groups, and because KCl and MgCl₂ protect the DNA from depurination and hydrolysis⁵⁴.

The G–C pair of nucleic acids is more thermostable than the A–T or A–U pairs because of the additional hydrogen bond⁵⁵. But elevated G + C ratios are not found among thermophilic prokaryotes because of the stability of the chromosomal DNA, although thermostability is correlated with G + C content of their ribosomal and transfer RNAs⁵⁶.

Low temperature

The fluidity of membranes decreases with decreasing temperature. In response, organisms increase the ratio of unsaturated to saturated fatty acids. In addition, the ability to withstand temperatures below freezing relies on two strategies: protection of the cells from ice formation by freezing avoidance, and if ice forms, protection from damage during thawing¹⁷. The proteins used in both processes are misleadingly named ‘antifreeze’ molecules — molecules that actually allow hysteresis to occur. In some terrestrial insects, hysteresis lowers the freezing point of water by 9–18 °C. Freezing of extracellular water during winter protects cells and is known from a small number of frogs, turtles and one snake⁵⁷.

Cold-temperature adaptation of protein occurs, although not always in ways that would be predicted from thermophile enzymes⁵⁸. At low temperatures there are low levels of free energy, so to decrease activation energy an enzyme must have a high degree of conformational complementarity with its substrate⁵⁹. At cold temperatures proteins become more rigid, implying that enhancing flexibility can restore function. Studies of α -amylase from the psychrophile *Alteromonas haloplanctis*, an enzyme with increased reliance of the molecular surface, a less rigid protein core and fewer interdomain interactions than mesophilic counterparts, have supported this hypothesis⁶⁰, as have studies of tubulin structure⁶¹.

Radiation and oxidative damage

Radiation and oxidative damage have always been common on Earth (L.J.R., C. L. Wilson, N. Chough and R. I. Donaldson, unpublished results). Mechanisms to avoid or repair environmentally encountered

Table 1 Classification and examples of extremophiles			
Environmental parameter	Type	Definition	Examples
Temperature	Hyperthermophile	Growth >80 °C	<i>Pyrolobus fumarii</i> , 113 °C
	Thermophile	Growth 60–80 °C	<i>Synechococcus lividis</i>
	Mesophile	15–60 °C	<i>Homo sapiens</i>
	Psychrophile	<15 °C	<i>Psychrobacter</i> , some insects
Radiation			<i>Deinococcus radiodurans</i>
Pressure	Barophile	Weight-loving	Unknown
	Piezophile	Pressure-loving	For microbe, 130 MPa
Gravity	Hypergravity	>1 g	None known
	Hypogravity	<1 g	None known
Vacuum		Tolerates vacuum (space devoid of matter)	Tardigrades, insects, microbes, seeds
Desiccation	Xerophiles	Anhydrobiotic	<i>Artemia salina</i> ; nematodes, microbes, fungi, lichens
Salinity	Halophile	Salt-loving (2–5 M NaCl)	Halobacteriaceae, <i>Dunaliella salina</i>
pH	Alkaliphile	pH > 9	<i>Natronobacterium</i> , <i>Bacillus firmus</i> OF4, <i>Spirulina</i> spp. (all pH 10.5)
	Acidophile	low pH-loving	<i>Cyanidium caldarium</i> , <i>Ferroplasma</i> sp. (both pH 0)
Oxygen tension	Anaerobe	Cannot tolerate O ₂	<i>Methanococcus jannaschii</i>
	Microaerophile	Tolerates some O ₂	<i>Clostridium</i>
Chemical extremes	Aerobe	Requires O ₂	<i>H. sapiens</i>
	Gases		<i>C. caldarium</i> (pure CO ₂)
	Metals	Can tolerate high concentrations of metal (metalotolerant)	<i>Ferroplasma acidarmanus</i> (Cu, As, Cd, Zn); <i>Ralstonia</i> sp. CH34 (Zn, Co, Cd, Hg, Pb)

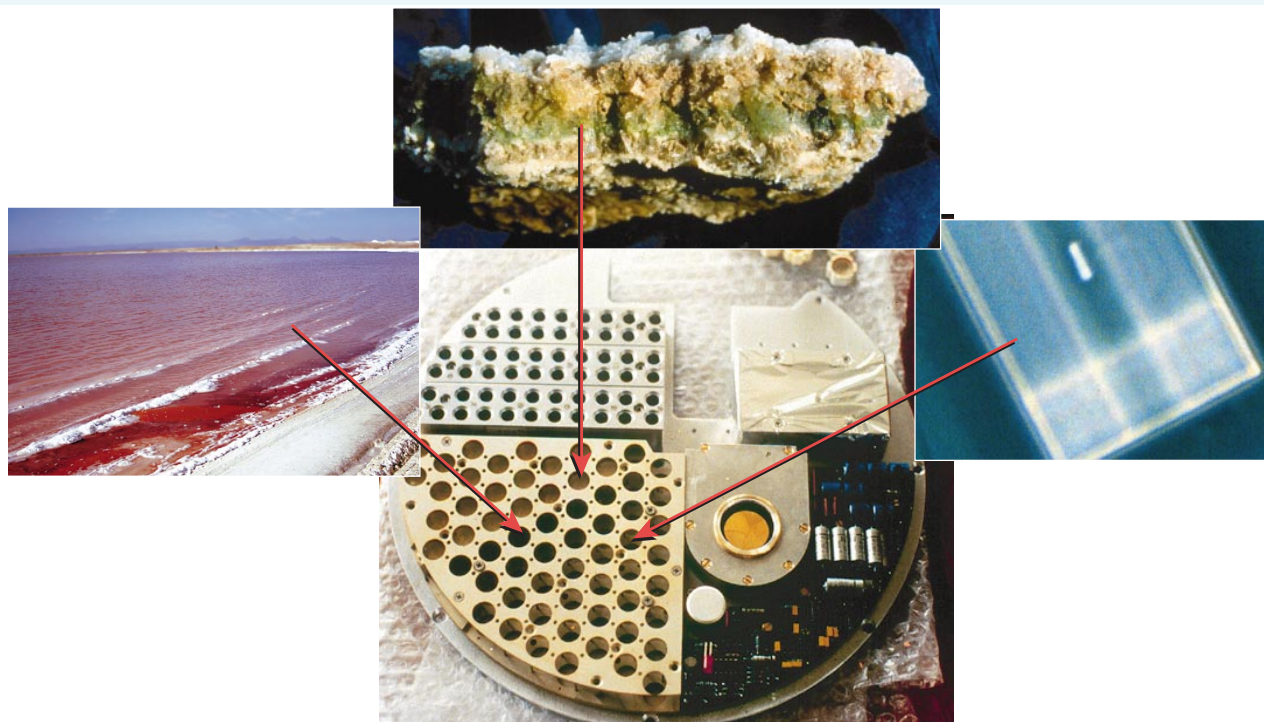


Figure 6 The BioPan halophile experiment. The BioPan facility was used to expose isolates of halophiles to the space environment in Earth's orbit during two two-week missions. Centred around the photograph of the internal sample-containing portion of the BioPan space hardware are, from left to right, a salt evaporation pond that appears red from the red-pigmented archaeal halophiles and some *Dunaliella*, an evaporite containing the cyanobacterium *Synechococcus* (Nägeli) collected from the Pacific marine intertidal zone, and a photomicrograph of a species of the extreme halophile *Halorcula* in a NaCl crystal.

damage include production of antioxidants and detoxifying enzymes, avoidance behaviour and repair mechanisms⁶². *D. radiodurans* copes with extraordinary radiation levels by containing a unique repair mechanism that involves reassembling of fragmented DNA^{21,63}.

Pressure

Pressure is known to alter gene expression⁶⁴. When pressure increases or temperature decreases, the molecules in lipid membranes pack tighter, resulting in decreased membrane fluidity²⁴. Often organisms circumvent this problem by increasing the proportion of unsaturated fatty acids in their membranes²⁵. Pressure can also help stabilize enzymes²⁴. High pressure can damage DNA and proteins in particular⁶⁵, so survival necessitates avoidance of damage or high repair rates.

Salinity and desiccation

Many microorganisms respond to increases in osmolarity by accumulating osmotica in their cytosol, which protects them from cytoplasmic dehydration and desiccation⁶⁶. With the exception of the Halobacteriaceae, which use K⁺ as their osmoticum⁶⁷, glycine betaine is the most effective osmoticum in most prokaryotes⁶⁸.

Osmotic concentration increases during desiccation, so responses are similar to those of a cell in high-salt environments. Compatible solutes such as K⁺, glutamate, glutamine, proline, glycine betaine, sucrose and trehalose accumulate away from proteins, forcing water nearby and thus stabilizing them³², and possibly stabilizing dry membranes⁶⁹. DNA damage is caused by increasing levels of desiccation from vacuum^{70,71}.

pH

Organisms that live at the extremes of pH are able to do so by maintaining their cytoplasm at the same pH as their mesophilic relatives, thus obviating the need for evolution of altered internal physiology. Active mechanisms to achieve this may involve secondary proton uptake mediated by membrane-associated antiporters. Passive mechanisms

include negatively charged cell-wall polymers in alkaliphiles⁴², and unusual bioenergetics, unusual permeability properties, positive surface charges, high internal buffer capacity, overexpression of H⁺ export enzymes and unique transporters for acidophiles³⁸.

Examples of extreme environment ecosystems

Hotsprings and geysers

Hotsprings and geysers are characterized by hot water and steam, and sometimes low pH and noxious elements such as mercury. The field was reviewed by Brock⁷², and much recent work⁷³ has been inspired by evolutionary biologists, biotechnology potential and astrobiology.

Deep sea

The deep-sea environment has high pressure and cold temperatures (1–2 °C), except in the vicinity of hydrothermal vents which are underwater geysers. In vents the temperature may be as high as 400 °C (ref. 74), but water remains liquid owing to the high hydrostatic pressure. Hydrothermal vents have a pH range from about 3 to 8 (ref. 75) and unusual chemistry²⁶. In 1977 the submarine *Alvin* found life 2.6 km deep along the East Pacific Rise, a centre of sea-floor spreading. Life forms range from microbes⁷⁶ to invertebrates²⁶.

Hydrothermal vents possibly were critical to evolution. Solution chemistry of hydrothermal vent systems is compatible with prebiotic chemistry leading to the origin of life⁷⁷ (but see ref. 55). Phylogenetic evidence points to thermophiles as the last common ancestor⁷⁸. Either life arose in a vent, or only thermophiles were able to survive the last of the major impacts during the late bombardment period⁷⁹.

Hypersaline environments

Hypersaline environments include salt flats, evaporation ponds, natural lakes (for example, Great Salt Lake) and deep-sea hypersaline basins⁴³. These communities often are dominated by halophilic archaea, including square archaea⁸⁰, or *D. salina*. Other organisms are

Figure 7 Mushroom Spring, Yellowstone National Park, USA, where Tom Brock isolated *Thermus aquaticus*, the organism from which Taq polymerase was obtained.



found at 25–33% salinity, including bacteria⁸¹ (for example, *Ectothiorhodospira halochloris*), cyanobacteria (for example, *Aphanothecae halophytica*, *Phormidium* sp. and *Schizothrix arenaria*), green algae (for example, *D. salina* and *Asteromonas gracilis*), diatoms (for example, *Amphora coffeaeformis* and species of *Navicula* and *Nitzschia*) and protozoa (for example, *Blepharisma halophila* and species of *Bodo*, *Phyllomitus* and *Tetramites*). There are halophilic yeasts and other fungi, but they are not nearly as halophilic as other microbial taxa.

Evaporites

Evaporite deposits consisting primarily of halite (NaCl), gypsum ($\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$) or anhydrite (CaSO_4) and containing bacterial and algal assemblages are well known in the fossil record⁸² and are still geographically widespread⁸³. Norton and Grant⁸⁴ showed that microorganisms entrapped in fluid inclusions of growing NaCl crystals may be motile for three weeks, and may remain viable for up to six months. Rothschild and colleagues⁸⁵ demonstrated that microorganisms inhabiting gypsum halite crusts perform carbon and nitrogen fixation while inside the dry crystals of the crust for at least a year. Although highly controversial, bacteria might survive for millions of years in the fluid inclusions of salt deposits including evaporites⁸⁶.

Deserts

Deserts are extremely dry, and cold or hot. Water is always a limiting factor, so such ecosystems are often dominated by microbiotic crusts⁸⁷. The Atacama Desert is one of the oldest, driest hot deserts on Earth⁸⁸. The coldest, driest places on Earth are the dry valleys of Antarctica. The primary inhabitants for both hot and cold deserts are cyanobacteria, algae and fungi that live a few millimetres beneath the sandstone rock surface. Although the endolithic communities in the Antarctic desert are based on photosynthesis (cyanobacteria, lichens and green algae⁸⁹), these microbes have adapted to long periods of darkness and dry conditions interspersed with dustings of dry snow, that upon melting are brief sources of water⁹⁰.

Ice, permafrost and snow

From high-altitude glaciers coloured pink with 'watermelon' or 'blood' snow (often green algae with photoprotective secondary carotenoids⁹¹) to the polar permafrost, microbial life has used frozen water as a habitat. But two caveats should be noted. First, some ice contains liquid brine inclusions that provide the actual habitat for the microbes⁹². Second, some ice environments such as permafrost contain "a community of survivors"⁹³. It is unlikely that the inhabitants of such an environment actually prefer this environment, rather they have found themselves trapped in the ice and are more

resistant than others that have suffered a similar fate. Microbial communities in sea ice contain algae (mostly diatoms), protozoa, bacteria and some archaea⁹⁴.

Atmosphere

The ability of an organism to survive in the atmosphere is a function of its ability to withstand desiccation and exposure to UV radiation^{95–97}. An airborne biota exists⁹⁸, although it is unclear whether it constitutes a functional ecosystem or is merely a live, but inactive, aerial suspension of organisms and their spore forms⁹⁹. Airborne organisms may travel across the Earth for hundreds to thousands of kilometres^{98,100}, and several kilometres up into the atmosphere¹⁰⁰. We argue that this field of aerobiology is critical to the enterprise of looking for life elsewhere in the Universe and further that it could be important in panspermia. In our view, it is one of the last frontiers of biological exploration on Earth, a view supported by the recent suggestion¹⁰¹ that life could have arisen in aerosols. On the present-day Earth, aerosols contain up to 50% organic material, and can acquire a lipid coating from the water below, meteorite-derived iron and nickel from the stratosphere, and energy from solar radiation — conditions conducive to the origin of life.

Space: new categories of extreme environments

Flight technology has enabled biological studies of space. Four main environments are currently of interest: manned-flight vehicles, interplanetary space (because of the potential for panspermia), and the planet Mars and jovian moon Europa (because of the possibility of liquid water and thus life) (Table 2). Thus, it is urgent that we define the environmental envelope for life, as well as conditions conducive to the origin of life, from hydrothermal to atmospheric¹⁰¹ to hypersaline¹⁰² parameters.

Mars

Mars is, for the most part, frigid (for current temperature, see <http://emma.la.asu.edu/daily.html>). The atmosphere receives 43% as much radiation as Earth, but attenuation through the thin,

Table 2 Physical conditions prevailing in interplanetary space

Parameter	Interplanetary space
Pressure (Pa)	10^{-14}
Solar electromagnetic radiation range	All
Cosmic ionizing radiation (Gy yr^{-1})	≤ 0.1
Gravity	$< 10^{-6}$ (varies*)
Temperature (K)	4 (varies*)

*Conditions vary depending on orientation and distance from the Sun.

Table 3 Examples of extremophiles in industry and biotechnology

Industrial process	Biomolecule	Advantages	Source organism
Hydrolysis of starch to produce soluble dextrins, maltodextrins and corn syrups	α -Amylase	High stability, aciduric, bacterial amylase	<i>Bacillus stearothermophilus</i> G-ZYME G995 (Enzyme Bio-System Ltd)
Paper bleaching	Xylanases	Decreases amount of bleach needed	Thermophiles
Prevent stalling in range of baked products	α -Amylase	Gives boost to yeast fermentation	Highest-stability bacterial amylase available, G-ZYME G-995
Food processing, baking, brewing, detergents	Proteases	Stable at high temperatures	Thermophiles
PCR reaction	DNA polymerase	No need to add additional enzyme during each cycle	Thermophiles
Cheese maturation, dairy production	Neutral proteases	Stable at low temperatures	Psychrophiles
Degradation of polymers in detergents	Proteases, amylases, lipases	Improved performance of detergent	Psychrophiles
Degradation of polymers in detergents	Cellulases, proteases, amylases, lipases	Stable at high pH	Alkaliphiles
Mariculture	Polyunsaturated fatty acids	Produced in cold temperatures	Psychrophiles
Bioremediation	Reduction of oil spills	Works efficiently in cold waters	Psychrophiles
Pharmaceuticals	Polyunsaturated fatty acids		Psychrophiles
Biosensors	Dehydrogenases		Psychrophiles
Desulphurization of coal	Sulphur oxidation		Acidophiles
Antibiotic production	Antibiotics		Alkaliphiles
Food colouring	Carotene	Inexpensive to produce	Halophiles/ <i>Dunaliella</i>
Pharmaceuticals	Glycerol, compatible solutes	Inexpensive to produce	Halophiles
Surfactants for pharmaceuticals	Membranes		Halophiles

CO₂-rich atmosphere is minimal, resulting in high surface fluxes of radiation >200 nm. Surface oxidants degrade organic carbon on the surface, which explains the negative results of the 1976 Viking missions¹⁰³. The atmospheric pressure is low (0.6–0.8 kPa), so liquid water is unstable on the surface, although hydrogeological evidence from the Mars Global Surveyor hints that liquid water may even flow today under the surface¹⁰⁴. Attention is now focused on the possibility of a subsurface biota, similar to the deep subsurface¹⁰⁵ or hydrothermal communities found on Earth.

Could life survive on the extreme harsh conditions of the martian surface? There are terrestrial organisms that hypothetically could withstand one or more of the martian extremes, but they would need protection¹⁰⁶. Mancinelli and Klovstad¹⁰⁷ demonstrated that a monolayer of *B. subtilis* spores protected by a 10- μ m-thick dust layer can survive UV exposure for weeks and probably years when exposed to a simulated martian UV-radiation flux. Thus, certain terrestrial microbes might survive on Mars.

Europa

Jupiter's moon Europa may harbour a subsurface water ocean. This putative ocean lies beneath an ice layer too thick to allow photosynthesis. However, Chyba has hypothesized¹⁰⁸ that disequilibrium chemistry in the ocean's ice cover, driven by charged particles accelerated in Jupiter's magnetosphere, could produce sufficient organic and oxidant molecules for a european biosphere. Lake Vostok in Antarctica possesses a perennially thick (3 km) ice-cover that precludes photosynthesis below, thus making it a good model

system for determining how a potential european biosphere might survive¹⁰⁹.

The space environment

The theory of panspermia, as proposed by Richter¹¹⁰, Lord Kelvin¹¹¹ and Arrhenius¹¹², holds that reproductive bodies of living organisms can exist throughout the Universe and develop wherever the environment is favourable. This implies that conditions favourable to the development of life prevailed at different locations in the Universe and at different times. Major criticisms of panspermia are that living organisms will not survive long exposure to space, and that it avoids the issue of where life began. But results of the Long Duration Exposure Facility and BioPan space experiments, which showed that microbes can survive in space, as well as the fact that organic compounds have been found in meteorites, has led to a re-examination of the feasibility of interplanetary transfer of living material, particularly microbes¹¹³.

Space is extremely cold, subject to unfiltered solar radiation, solar wind, galactic radiation, space vacuum and negligible gravity^{105,114}. At the distance of the Earth from the Sun, solar irradiance is 1,360 W m⁻². Of this, 45% is infrared light, 48% visible and only 7% UV. Space is a nutritional wasteland with respect to water and organic compounds, although comets may provide an oasis when passing a warming star.

Terrestrial organisms most likely to survive these conditions are microbes, with comets or meteorites as conveyance. Microgravity is not lethal; cold tolerance and anhydrobiosis are survivable. Until we understand transit times, we cannot address adequately the nutritional needs of organisms in transit, but we hypothesize that with the exceedingly low metabolic rates that would result from the extremes in cold and desiccation, nutritional needs would not exist. Thus, we are left with two potential 'show-stoppers': radiation and the space vacuum. Heavy ions are mutagenic or lethal to microbes¹¹⁵. Most damage to microbes exposed to space is due to UV radiation, especially during the short term, but heavy ionizing radiation has a greater probability of being lethal.

Remarkably, some terrestrial organisms can survive this highly extreme environment. This has been proven through flight experiments led by the European Space Agency with American participation (Fig. 6). Microbes tested in the space environment and then returned to Earth include *B. subtilis* spores, bacteriophage T-1, tobacco mosaic virus¹¹³, and most recently osmophilic microbes. *B. subtilis* spores will survive for years in space if either in a bilayer (or multilayer) or mixed with glucose to protect them against high solar UV-radiation flux, but if they are exposed in a monolayer they are killed within minutes¹¹³. For comparison, viruses lose viability by weeks. Although the data are controversial, *D. radiodurans* did not survive 7 months in space and the DNA had extensive breakage³⁴. Halophiles can survive for two weeks in space and probably much longer (R.L.M., M. R. Klovstad, P. Rettberg, & G. Horneck, unpublished results). The halophiles are the first example of a vegetative cell surviving exposure to the space environment.

Economic potential of extremophiles

Extremophiles have provided data that are basic to molecular biology, including information on protein folding. Evolutionary biology has benefited on two fronts. First, in the race to uncover the most extreme of extremophiles, whole new taxa have been discovered, increasing phylogenetic enlightenment. Second, the ability to survive in some extreme environments has evolved multiple times, leading to a new understanding of chance versus necessity in evolutionary pathways, especially at the molecular level. For example, the ice-binding antifreeze proteins are evolutionarily convergent, with that of the Antarctic notothenioid fish evolving from a pancreatic trypsinogen-like protease¹¹⁶.

Extremophiles have endeared themselves to multibillion-dollar industries, including agricultural, chemical synthesis, laundry

detergents¹¹⁷ and pharmaceuticals. The European Commission has supported research, training and the commercialization of technology in this area⁶ since 1982. From 1996–1999 it funded the ‘Extremophiles as Cell Factories’ project (see http://www.tuttech.de/ecf/ecf1_3.htm), which is now in a phase of industry-sponsored technology transfer to European companies (G. Antranikian, personal communication). Enzymes are sought that are stable and functional in economically preferable environments, such as high or unstable temperatures¹¹⁸ (Table 3).

Enzymes from extremophiles — ‘extremozymes’¹¹⁹ — have potential in multiple areas, either by using the enzymes themselves, or by using them as sources of ideas to modify mesophile-derived enzymes. In most cases the reaction medium is aqueous, although results have indicated that aqueous/organic and nonaqueous media allow the modification of reaction equilibria and enzyme specificity, creating pathways for synthesizing novel compounds¹²⁰. The fastidious growth conditions for extremophiles means that it is often economically advantageous to express the gene in a more tractable host organism such as *E. coli*.

The canonical example of extremophile-derived enzymes in biotechnology is the source of Taq polymerase, the enzyme at the crux of the widely used polymerase chain reaction (PCR). Taq polymerase was isolated from the thermophilic bacterium *Thermus aquaticus*, an organism discovered in 1969 in Yellowstone National Park, Wyoming (ref. 121, Fig. 7). DNA polymerases from other thermophiles have been marketed by Promega Corporation as a product for high-fidelity PCR, with each having its own advantages^{122,123}.

Other extremophiles have industrial applications. For example, some Antarctic bacteria produce polyunsaturated fatty acids, an essential dietary ingredient for many aquaculture species (for example, Atlantic salmon). The bacteria are used to enrich rotifers, a food organism for larval fish¹²⁴. Antarctic bacteria have potential in bioremediation of waters following oil spills, which is a concern in cold waters¹²⁴. *D. salina* is widely used for the commercial production of β -carotenes, which it produces in response to solar radiation, and glycerol, which it produces to counterbalance external osmotic pressure¹²⁵.

Human health may benefit from extremophiles indirectly through biotechnology and bioremediation (Table 3). Direct uses include marketing of dried *Dunaliella* as a nutritional supplement, primarily as an antioxidant. Antifreeze proteins show potential as cryoprotectants of frozen organs.

What next?

Extremophile research is entering an exciting phase. The commercial potential has been recognized, but is far from being realized. Our ignorance of microbial diversity coupled with improvements in exploration and analytical technology suggest that many more discoveries will be forthcoming. The International Space Station will enhance long-term biological studies in space, improving our understanding of the scope of that formerly inaccessible environment. Colonization and terraforming of Mars will require a supporting biota, and where better to start than with extremophiles? And, when life severs its links to planet Earth it will enter new niches ripe for extremophiles, perhaps joining indigenous extraterrestrial extremophiles. □

1. Macelroy, R. D. Some comments on the evolution of extremophiles. *Biosystems* **6**, 74–75 (1974).
2. Madigan, M. T. & Marrs, B. L. Extremophiles. *Sci. Am.* **276**, 82–87 (1997).
3. Horikoshi, K. & Grant, W. D. *Extremophiles. Microbial Life in Extreme Environments* (Wiley-Liss, New York, 1998).
4. Seckbach, J. (ed.) *Journey to Diverse Microbial Worlds: Adaptation to Exotic Environments* (Kluwer, Dordrecht, 2000).
5. Cowan, D. Hot bugs, cold bugs and sushi. *Trends Biotechnol.* **16**, 241–242 (1998).
6. Aguilar, A., Ingemansson, T. & Magnien, E. Extremophile microorganisms as cell factories: support from the European Union. *Extremophiles* **2**, 367–373 (1998).
7. Tyrell, R. M. in *Oxidative Stress: Oxidants and Antioxidants* (ed. Sies, H.) 57–83 (Academic, London, 1991).
8. Newcomb, T. G. & Loeb, L. A. in *DNA Damage and Repair, Vol. 1: DNA Repair in Prokaryotes and Lower Eukaryotes* (eds Nickoloff, J. A. & Hoekstra, M. F.) 65–84 (Humana, Totowa, NJ, 1998).

9. Minton, K. W. DNA repair in the extremely radioresistant bacterium *Deinococcus radiodurans*. *Mol. Microbiol.* **13**, 9–15 (1994).
10. Chow, F. I. & Tan, S. T. Manganese(II) induces cell division and increases in superoxide dismutase and catalase activities in an aging deinococcal culture. *J. Bacteriol.* **172**, 2029–2035 (1990).
11. Venkateswaran, A. et al. Physiologic determinants of radiation resistance in *Deinococcus radiodurans*. *Appl. Environ. Microbiol.* **66**, 2620–2626 (2000).
12. Seki, K. & Toyoshima, M. Preserving tardigrades under pressure. *Nature* **395**, 853–854 (1998).
13. Ball, P. *H₂O: A Biography of Water* (Weldonfeld & Nicolson, London, 1999).
14. Morita, R. Y. Psychrophilic bacteria. *Bacteriol. Rev.* **39**, 144–167 (1975).
15. Blochl, E. et al. *Pyrololobus fumarii*, gen. and sp. nov., represents a novel group of archaea, extending the upper temperature limit for life to 113 °C. *Extremophiles* **1**, 14–21 (1997).
16. Schuliger, J. W., Brown, S. H., Baross, J. A. & Kelly, R. M. Purification and characterization of a novel amylolytic enzyme from ES4, a marine hyperthermophilic archaeum. *Mol. Mar. Biol. Biotech.* **2**, 76–87 (1993).
17. Clarke, A. in *Evolution on Planet Earth: The Impact of the Physical Environment* (eds Rothschild, L. & Lister, A.) (Academic, London, in the press).
18. Kohshima, S. A novel cold-tolerant insect found in a Himalayan glacier. *Nature* **310**, 225 (1984).
19. Franks, F. *Biophysics and Biochemistry at Low Temperatures* (Cambridge Univ. Press, Cambridge, 1985).
20. Wharton, D. A. & Ferns, D. J. Survival of intracellular freezing by the Antarctic nematode *Panagrolaimus davidi*. *J. Exp. Biol.* **198**, 1381–1387 (1995).
21. Battista, J. R. Against all odds: the survival strategies of *Deinococcus radiodurans*. *Annu. Rev. Microbiol.* **51**, 203–224 (1997).
22. Ferreira, A. C. et al. Characterization and radiation resistance of new isolates of *Rubrobacter radiotolerans* and *Rubrobacter xylanophilus*. *Extremophiles* **3**, 235–238 (1999).
23. Ben-Amotz, A. & Avron, M. *Dunaliella bardawil* can survive especially high irradiance levels by the accumulation of β -carotene. *Trends Biotechnol.* **8**, 121–126 (1990).
24. Pledger, R. J., Crump, B. C. & Baross, J. A. A barophilic response by two hyperthermophilic, hydrothermal vent *Archaea*: an upward shift in the optimal temperature and acceleration of growth rate at supra-optimal temperatures by elevated pressure. *FEMS Microbiol. Ecol.* **14**, 233–242 (1994).
25. Bartlett, D. H. & Bidle, K. A. in *Enigmatic Microorganisms and Life in Extreme Environments* (ed. Seckbach, J.) 503–512 (Kluwer, Dordrecht, 1999).
26. Van Dover, C. L. *The Ecology of Deep-Sea Hydrothermal Vents* (Princeton Univ. Press, Princeton, 2000).
27. Kato, C. et al. Extremely barophilic bacteria isolated from the Mariana Trench, Challenger Deep, at a depth of 11,000 meters. *Appl. Environ. Microbiol.* **64**, 1510–1513 (1998).
28. Cogoli, A., Iversen, T. H., Johnsson, A., Mesland, D. & Oser, H. European Space Agency Spec. Publ. No. 1105, 49–64 (1989).
29. Crowe, J. H. Anhydrobiosis: an unsolved problem. *Am. Nat.* **105**, 563–574 (1971).
30. Wright, J. C. Desiccation tolerance and water-retentive mechanisms in tardigrades. *J. Exp. Biol.* **142**, 267–292 (1989).
31. Glasheen, J. S. & Hand, S. C. Anhydrobiosis in embryos of the brine shrimp *Artemia*: characterization of metabolic arrest during reductions in cell-associated water. *J. Exp. Biol.* **135**, 363–389 (1988).
32. Potts, M. Desiccation tolerance of prokaryotes. *Microbiol. Rev.* **58**, 755–805 (1994).
33. Cox, C. S. Roles of water molecules in bacteria and viruses. *Origins Life* **23**, 29–36 (1993).
34. Dose, K. et al. ERA-experiment: space biochemistry. *Adv. Space Res.* **16**(8), 119–129 (1995).
35. Dose, K. & Gill, M. DNA stability and survival of *Bacillus subtilis* spores in extreme dryness. *Origins Life* **25**, 277–293 (1994).
36. Seckbach, J. in *Enigmatic Microorganisms and Life in Extreme Environments* (ed. Seckbach, J.) 427–435 (Kluwer, Dordrecht, 1999).
37. Doemel, W. N. & Brock, T. D. The physiological ecology of *Cyanidium caldarium*. *J. Gen. Microbiol.* **67**, 17–32 (1971).
38. Pick, U. in *Enigmatic Microorganisms and Life in Extreme Environments* (ed. Seckbach, J.) 467–478 (Kluwer, Dordrecht, 1999).
39. Schleper, C., Pühler, G., Kühlmorgen, B. & Zillig, W. Life at extremely low pH. *Nature* **375**, 741–742 (1995).
40. Schleper, C. et al. *Picrophilus* gen. nov., fam. nov.: a novel aerobic, heterotrophic, thermoadaptophilic genus and family comprising archaea capable of growth around pH 0. *J. Bacteriol.* **177**, 7050–7059 (1995).
41. Edwards, K. J., Bond, P. L., Gihring, T. M. & Banfield, J. F. An archaeal iron-oxidizing extreme acidophile important in acid mine drainage. *Science* **287**, 1796–1799 (2000).
42. Krulwich, T. A., Ito, M., Hicks, D. B., Gilmour, R. & Guffanti, A. A pH homeostasis and ATP synthesis: studies of two processes that necessitate inward proton translocation in extremely alkaliphilic *Bacillus* species. *Extremophiles* **2**, 217–222 (1998).
43. Javor, B. *Hypersaline Environments* (Springer, Berlin, 1989).
44. Jones, B. E., Grant, W. D., Duckworth, A. W. & Owensson, G. G. Microbial diversity of soda lakes. *Extremophiles* **2**, 191–200 (1998).
45. Beckman, K. B. & Ames, B. N. The free radical theory of aging matures. *Physiol. Rev.* **78**, 547–581 (1998).
46. Pourzand, C. & Tyrrell, R. M. Apoptosis, the role of oxidative stress and the example of solar UV radiation. *Photochem. Photobiol.* **70**, 380–390 (1999).
47. Cooper, W. & Lean, D. in *Encyclopedia of Earth System Science* Vol. 2 (ed. Nierenber, W. A.) 527–535 (Academic, San Diego, 1992).
48. Seckbach, J., Baker, F. A. & Shugarman, P. M. Algae survive under pure CO₂. *Nature* **227**, 744–745 (1970).
49. Nies, D. H. Heavy metal-resistant bacteria as extremophiles: molecular physiology and biotechnological use of *Ralstonia* sp. CH34. *Extremophiles* **4**, 77–82 (2000).
50. Isken, S. & de Bont, J. A. M. Bacteria tolerant to organic solvents. *Extremophiles* **2**, 229–238 (1998).
51. Beardall, J. & Entwistle, L. Internal pH of the obligate acidophile *Cyanidium caldarium* Geitler (Rhodophyta?). *Phycologia* **23**, 397–399 (1984).
52. Jaenicke, R. Stability and folding of ultrastable proteins: eye lens crystallins and enzymes from thermophiles. *FASEB J.* **10**, 84–92 (1996).
53. Peak, M. J., Robb, F. T. & Peak, J. G. Extreme resistance to thermally induced DNA backbone breaks in the hyperthermophilic archaeon *Pyrococcus furiosus*. *J. Bacteriol.* **177**, 6316–6318 (1995).
54. Marguet, E. & Forterre, P. Protection of DNA by salts against thermodegradation at temperatures typical for hyperthermophiles. *Extremophiles* **2**, 115–122 (1998).

55. Galtier, N., Tourasse, N. & Gouy, M. A nonhyperthermophilic common ancestor to extant life forms. *Science* **283**, 220–221 (1999).
56. Galtier, N. & Lobry, J. R. Relationships between genomic G + C content, secondary structures and optimal growth temperature in prokaryotes. *J. Mol. Evol.* **44**, 632 (1997).
57. Storey, K. B. & Storey, J. M. Natural freezing survival in animals. *Annu. Rev. Ecol. Syst.* **27**, 365–386 (1996).
58. Russell, N. J. Toward a molecular understanding of cold activity of enzymes from psychrophiles. *Extremophiles* **4**, 83–90 (2000).
59. Cummings, S. P. & Black, G. W. Polymer hydrolysis in a cold climate. *Extremophiles* **3**, 81–87 (1999).
60. Aghajari, N., Feller, G., Gerday, C. & Haser, R. Structures of the psychrophilic *Alteromonas haloplantici* α -amylase give insights into cold adaptation at a molecular level. *Structure* **6**, 1503–1516 (1998).
61. Willem, S. *et al.* Protein adaptation to low temperatures: a comparative study of α -tubulin sequences in mesophilic and psychrophilic algae. *Extremophiles* **3**, 221–226 (1999).
62. Rothschild, L. J. in *Enigmatic Microorganisms and Life in Extreme Environments* (ed. Seckbach, J.) 551–562 (Kluwer, Dordrecht, 1999).
63. Battista, J. R. in *DNA Damage and Repair, Vol. I: DNA Repair in Prokaryotes and Lower Eukaryotes* (eds Nickoloff, J. A. & Hoekstra, M. F.) 287–303 (Humana, Totowa, NJ, 1998).
64. Nakasone, K., Ikegami, A., Kato, C., Usami, R. & Horikoshi, K. Mechanisms of gene expression controlled by pressure in deep-sea microorganisms. *Extremophiles* **2**, 149–154 (1998).
65. Abe, F., Kato, C. & Horikoshi, K. Pressure-regulated metabolism in microorganisms. *Trends Microbiol.* **7**, 447–453 (1999).
66. Yancey, P. H., Clark, M. E., Hand, S. C., Bowlus, R. D. & Somero, G. N. Living with water stress: evolution of osmolyte systems. *Science* **217**, 1214–1216 (1982).
67. Larsen, H. Biochemical aspects of extreme halophilism. *Adv. Microb. Physiol.* **1**, 97–132 (1967).
68. Le Rudulier, D. & Bouillard, L. Glycine Betaine, an osmotic effector in *Klebsiella pneumoniae* and other members of the Enterobacteriaceae. *Appl. Environ. Microbiol.* **46**, 152–159 (1983).
69. Crowe, J. H., Hoekstra, F. A. & Crowe, L. M. Anhydrobiosis. *Annu. Rev. Physiol.* **54**, 579–599 (1992).
70. Wehner, J. & Horneck, G. Effects of vacuum UV and UVC radiation on dry *E. coli* plasmid pUC19 II. Mutational specificity at the *lacZ* gene. *J. Photochem. Photobiol. B* **30**, 171–177 (1995).
71. Wehner, J. & Horneck, G. Effects of vacuum UV and UVC radiation on dry *E. coli* plasmid pUC19 I. Inactivation, *lacZ* mutation induction and strand breaks. *J. Photochem. Photobiol. B* **28**, 77–85 (1995).
72. Brock, T. D. *Thermophilic Microorganisms and Life at High Temperatures* (Springer, New York, 1978).
73. Reysenbach, A. L., Voytek, M. & Mancinelli, R. L. (eds) *Microbiology of Yellowstone* (Kluwer, New York, in the press).
74. Horikoshi, K. Barophiles: deep-sea microorganisms adapted to an extreme environment. *Curr. Opin. Microbiol.* **1**, 291–295 (1998).
75. Kennish, M. J. (ed.) *Practical Handbook of Marine Science* 2nd edn 236–237 (CRC Press, Boca Raton, 1994).
76. Karl, D. M. (ed.) *The Microbiology of Deep-sea Hydrothermal Vents* (CRC Press, Boca Raton, 1995).
77. Cody, G. D. *et al.* Primordial carbonylated iron-sulfur compounds and the synthesis of pyruvate. *Science* **289**, 1337–1340 (2000).
78. Pace, N. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).
79. Sleep, N. H., Zahnle, K. J., Kasting, J. F. & Morowitz, H. J. Annihilation of ecosystems by large impacts on early earth. *Nature* **342**, 139–142 (1989).
80. Oren, A. in *Enigmatic Microorganisms and Life in Extreme Environments* (ed. Seckbach, J.) 339–355 (Kluwer, Dordrecht, 1999).
81. Kamekura, M. Diversity of extremely halophilic bacteria. *Extremophiles* **2**, 289–295 (1998).
82. Bell, C. M. Saline lake carbonates within an Upper Jurassic-Lower Cretaceous continental red bed sequence in the Atacama region of northern Chile. *Sedimentology* **36**, 651–664 (1989).
83. Castanier, S., Perthuisot, J.-P., Rouchy, J.-M., Maurin, A. & Guelorget, O. Halite ooids in Lake Asal Djibouti biocrystalline build-ups. *Geobios (Lyon)* **25**, 811–821 (1992).
84. Norton, C. F. & Grant, W. D. Survival of halobacteria within fluid inclusions in salt crystals. *J. Gen. Microbiol.* **134**, 1365–1373 (1988).
85. Rothschild, L. J., Giver, L. J., White, M. R. & Mancinelli, R. L. Metabolic activity of microorganisms in gypsum-halite crusts. *J. Phycol.* **30**, 431–438 (1994).
86. Vreeland, R. H., Rosenzweig, W. D. & Powers, D. W. Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal. *Nature* **407**, 897–900 (2000).
87. Evans, R. D. & Johansen, J. R. Microbiotic crusts and ecosystem processes. *Crit. Rev. Plant Sci.* **18**, 182–225 (1999).
88. Rundel, P. W. *et al.* The phytogeography and ecology of the coastal Atacama and Peruvian deserts. *ALISO* **13**, 1–49 (1991).
89. van Thiel, N. & Garbary, D. J. in *Enigmatic Microorganisms and Life in Extreme Environments* (ed. Seckbach, J.) 245–253 (Kluwer, Dordrecht, 1999).
90. Friedmann, E. I. Endolithic microorganisms in the Antarctic cold desert. *Science* **215**, 1045–1053 (1982).
91. Bidigare, R. R. *et al.* Evidence for a photoprotective function for secondary carotenoids of snow algae. *J. Phycol.* **29**, 427–434 (1993).
92. Junge, K., Krembs, C., Deming, J., Stierle, A. & Eicken, H. A microscopic approach to investigate bacteria under in-situ conditions in sea-ice samples. *Ann. Glaciol.* (in the press).
93. Friedmann, E. I. *Viable Microorganisms in Permafrost* (ed. Gilichinsky, D. A.) 21–26 (Institute of Soil Science and Photosynthesis, Russian Academy of Science, Pushchino, 1994); cited in Vishnivetskaya, T., Kathariou, S., McGrath, J., Gilichinsky, D. & Tiedje, J. M. Low-temperature recovery strategies for the isolation of bacteria from ancient permafrost sediments. *Extremophiles* **4**, 165–173 (2000).
94. Staley, J. T. & Gosink, J. J. Poles apart: biodiversity and biogeography of sea ice bacteria. *Annu. Rev. Microbiol.* **53**, 189–215 (1999).
95. Mancinelli, R. L. & Shulls, W. A. Airborne bacteria in an urban environment. *Appl. Environ. Microbiol.* **35**, 1095–1101 (1978).
96. Cox, C. S. *Roles of Maillard Reactions in Diseases* (HMSO, London, 1991).
97. Israeli, E., Gitelman, J. & Lighthart, B. in *Atmospheric Microbial Aerosol Theory and Application* (eds Lighthart, B. & Mohr, A. J.) 166–192 (Chapman & Hall, New York, 1994).
98. Cox, C. S. *The Aerobiological Pathway of Microorganisms* (Wiley, New York, 1987).
99. Cox, C. S. & Wathes, C. M. *Bioaerosols Handbook* (Lewis, London, 1995).
100. Lighthart, B. & Mohr, A. J. (eds) *Atmospheric Microbial Aerosol Theory and Application* 68–98, 166–192 (Chapman & Hall, New York, 1994).
101. Marchant, J. Life from the skies—did droplets high in the atmosphere give birth to the first living cells? *New Sci.* **2247**, 4–5 (15 July 2000).
102. Dundas, I. Was the environment for primordial life hypersaline? *Extremophiles* **2**, 375–377 (1998).
103. Biemann, K. *et al.* The search for organic substances and inorganic volatile compounds in the surface of Mars. *J. Geophys. Res.* **82**, 4641–4658 (1977).
104. Malin, M. C. & Edgett, K. S. Evidence for recent groundwater seepage and surface runoff on Mars. *Science* **288**, 2330–2335 (2000).
105. Horneck, G. in *Evolution on Planet Earth: The Impact of the Physical Environment* (eds Rothschild, L. & Lister, A.) (Academic, London, in the press).
106. Rothschild, L. J. Earth analogs for Martian life. Microbes in evaporites, a new model system for life on Mars. *Icarus* **88**, 246–260 (1990).
107. Mancinelli, R. L. & Klovstad, M. Survival of *Bacillus subtilis* spores on space craft surfaces. *Planet. Space Sci.* **48**, 1093–1097 (2000).
108. Chyba, C. Energy for microbial life on Europa. *Nature* **403**, 381–382 (2000).
109. Stone, R. Permafrost comes alive for Siberian researchers. *Science* **286**, 36–37 (1999).
110. Richter, H. *Zur Darwinschen Lehre. Schmidts Jahrb. Ges. Med.* **126**, 243–249 (1865).
111. Thomson, W. in *Popular Lectures and Addresses* 132–205 (Macmillan, New York, 1894).
112. Arrhenius, S. Die Verbreitung des Lebens im Weltenraum. *Umschau* **7**, 481–485 (1903).
113. Nicholson, W. L., Munakata, N., Horneck, G., Melosh, H. J. & Setlow, P. Resistance of *Bacillus* endospores to extreme terrestrial and extraterrestrial environments. *Microbiol. Mol. Biol. Rev.* **64**, 548–572 (2000).
114. Rothschild, L. J. in *Evolution on Planet Earth: The Impact of the Physical Environment* (eds Rothschild, L. & Lister, A.) (Academic, London, in the press).
115. Horneck, G., Krasavin, E. A. & Kozubek, S. Mutagenic effects of heavy ions in bacteria. *Adv. Space Res.* **14**(10), 315–329 (1994).
116. Cheng, C.-H. C. & Chen, L. Evolution of an antifreeze glycoprotein. *Nature* **401**, 463–464 (1999).
117. Ito, S. *et al.* Alkaline detergent enzymes from alkaliphiles: enzymatic properties, genetics, and structures. *Extremophiles* **2**, 185–190 (1998).
118. Zeikus, J. G., Vielle, C. & Savchenko, A. Thermozymes: biotechnology and structure-function relationships. *Extremophiles* **2**, 179–182 (1998).
119. Hough, D. W. & Danson, M. J. Extremozymes. *Curr. Opin. Chem. Biol.* **3**, 39–46 (1999).
120. Sellek, G. A. & Chaudhuri, J. B. Biocatalysis in organic media using enzymes from extremophiles. *Enzyme Microb. Technol.* **25**, 471–482 (1999).
121. Brock, T. D. & Freeze, H. *Thermus aquaticus* gen. n., a nonsporulating extreme thermophile. *J. Bacteriol.* **98**, 289–297 (1969).
122. Mattila, P., Korpela, J., Tenkanen, T. & Pitkanen, K. Fidelity of DNA synthesis by the *Thermococcus litoralis* DNA polymerase—an extremely heat stable enzyme with proofreading activity. *Nucleic Acids Res.* **19**, 4967–4973 (1991).
123. Cariello, N. F., Swenberg, J. A. & Skopek, T. R. Fidelity of *Thermococcus litoralis* DNA polymerase (Vent) in PCR determined by denaturing gradient gel electrophoresis. *Nucleic Acids Res.* **19**, 4193–4198 (1991).
124. Nichols, D. *et al.* Developments with Antarctic microorganisms: culture collections, bioactivity screening, taxonomy, PUFA production and cold-adapted enzymes. *Curr. Opin. Biotechnol.* **10**, 240–246 (1999).
125. Ben-Amotz, A. in *Enigmatic Microorganisms and Life in Extreme Environments* (ed. Seckbach, J.) 401–410 (Kluwer, Dordrecht, 1999).

Acknowledgements

We thank the many people who were generous with information, especially: J. Baross on hydrothermal vents; L. Giver and C. Wong on commercial aspects; G. Antranikian and M. Meyer on government programmes; J. Deming, K. Junge, P. Ball, S. Emerson and G. Packard on life at low temperatures; and K. Stedman for life at high temperatures. A. Deutch, K. Duffy and S. Sturtevant provided tips on the thermophiles of Yellowstone. E. Holton, D. Cowan and J. Parkes provided helpful reviews.

Chance and necessity: the evolution of morphological complexity and diversity

Sean B. Carroll

Howard Hughes Medical Institute and Laboratory of Molecular Biology, University of Wisconsin–Madison, 1525 Linden Drive, Madison, Wisconsin 53706-1596, USA (e-mail sbcarroll@facstaff.wisc.edu)

The primary foundation for contemplating the possible forms of life elsewhere in the Universe is the evolutionary trends that have marked life on Earth. For its first three billion years, life on Earth was a world of microscopic forms, rarely achieving a size greater than a millimetre or a complexity beyond two or three cell types. But in the past 600 million years, the evolution of much larger and more complex organisms has transformed the biosphere. Despite their disparate forms and physiologies, the evolution and diversification of plants, animals, fungi and other macroforms has followed similar global trends. One of the most important features underlying evolutionary increases in animal and plant size, complexity and diversity has been their modular construction from reiterated parts. Although simple filamentous and spherical forms may evolve wherever cellular life exists, the evolution of motile, modular mega-organisms might not be a universal pattern.

“Drawn out of the realm of pure chance, the accident enters into that of necessity, of the most implacable certainties.”
J. Monod¹

It is widely accepted that the evolution of any particular organism or form is a product of the interplay of a great number of historical contingencies¹. Rewind and replay the tape of life again and again, as the now familiar argument goes, and there is no predicting (or reproducing) the outcomes. Roses and redwoods, humans and hummingbirds, trilobites and dinosaurs each owe their existence (or demise) to unfathomable combinations of innumerable rolls of the ecological and genetic dice.

Life's contingent history could be viewed as an argument against any direction or pattern in the course of evolution or the shape of life. But it is obvious that larger and more complex life forms have evolved from simple unicellular ancestors and that various innovations were necessary for the evolution of new means of living. This raises the possibility that there are trends within evolutionary history that might reflect the existence of general principles governing the evolution of increasingly larger and more complex forms. The first task of this review is to examine the degree to which the evolution of the shapes of life are a matter of chance — a random walk in morphospace — or of necessity — borne from the demands of natural selection and the constraints imposed by physics, genetics and development. The second task is to extrapolate from the evolutionary trends on Earth to assess what they might portend for the evolution of life elsewhere.

There is a long history of support for the general notion of overall evolutionary trends towards increases in size², complexity^{2,3} and diversity^{4–6}. However, there are two fundamentally distinct mechanisms that have been proposed to explain these trends⁷. One is a random, passive tendency to evolve away from the initial minima of organismal size⁸, complexity and diversity through an overall increase in variance (that is, there is “nowhere to go but up”)^{9,10}. The second is a non-random, active or ‘driven’ process that biases evolution towards increased size or complexity¹¹.

There are relationships between size and complexity and between complexity and diversity that are intuitive. Increases in organismal size through increases in cell number create the potential for increases in diversity of cell type and, as a result,

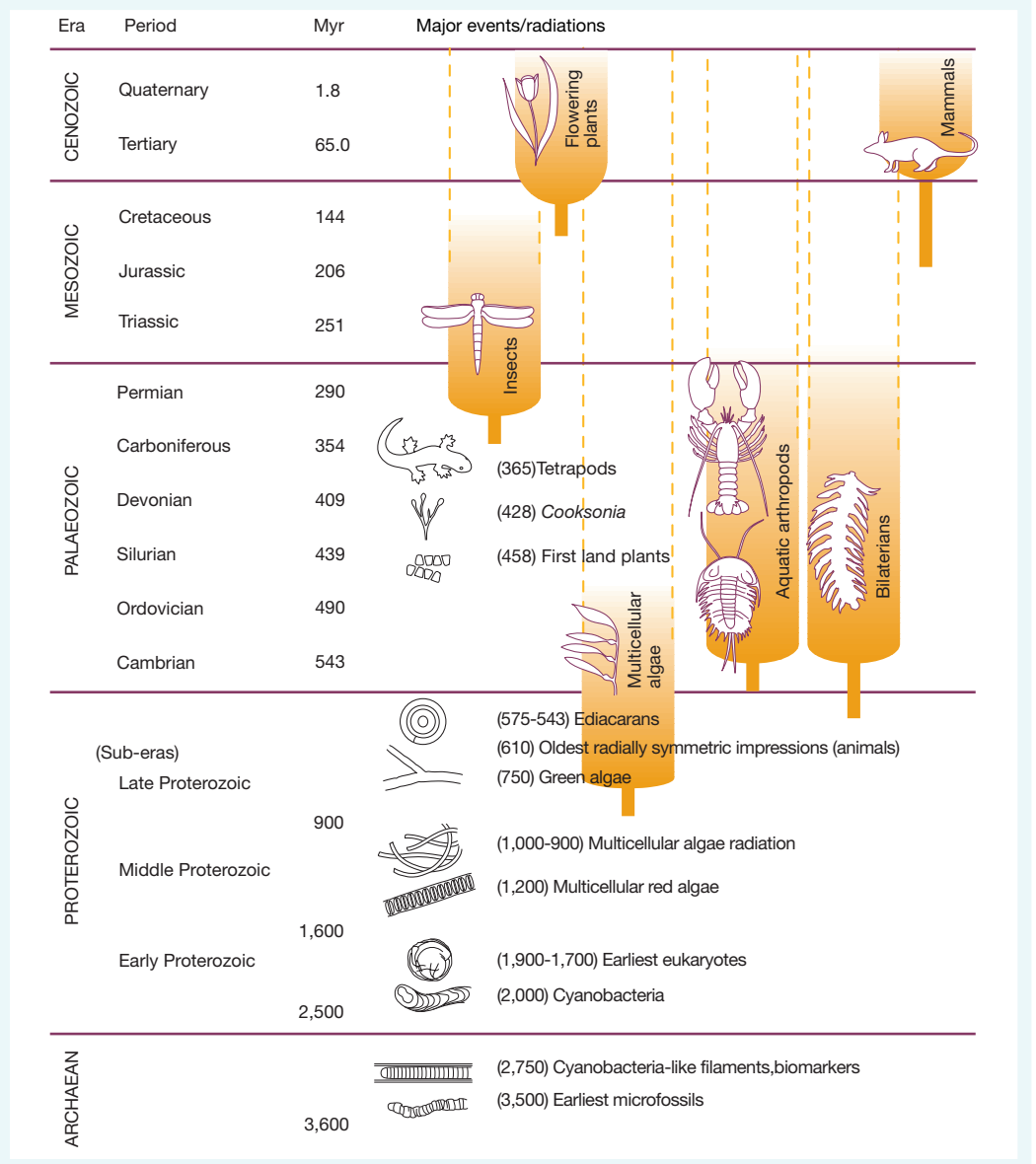
anatomical complexity. Increases in morphological complexity may lead to expansions into previously unoccupied ‘ecospace’ and accompanying expansions of species diversity.

Testing these relationships, searching for trends, and identifying potential causes requires analyses far beyond the mere description of the global history from bacteria to whales, to the consideration of detailed histories of specific lineages. Singularities are dangerous territory for the formulation of general trends and principles¹²; fortunately, however, aspects of evolutionary history have been repeated in different lineages. Here, I will focus on events that have occurred and trends that are manifest in a variety of macroforms with the aim of identifying some of the trends and potential underlying forces that have shaped the size, complexity and diversity of macroscopic life. I will first review the evidence for global trends and then consider the histories of particular lineages where the mechanisms underlying the generation of morphological complexity and the evolution of diversity are beginning to be better understood. In particular, I will focus on the relationship between genomic and organismal complexity that can now be addressed by recent advances in the analysis of genes, genomes and development. I will develop the argument that one of the most important features that has facilitated the evolution of plant and animal complexity and diversity is the modularity of their construction from reiterated, differentiated parts. Finally, I will discuss which trends in the evolution of morphology are likely to apply wherever life may be found.

Milestones in morphological evolution

The principal events of interest here are the major changes in organismal size, form and complexity, and the major expansions in diversity, that have produced the many shapes of macroscopic life. The foundations for inferences about the sequence and direction of evolution are the fossil record and the phylogenetic tree of life. Integration of palaeontological and systematic data is required to establish the number of times particular events occurred, the order in which important sets of traits evolved, and to identify the possible sister groups of major taxonomic groups. The fossil record is also a primary source of data on the time of origin of taxa. One must

Figure 1 History of major evolutionary events from the fossil record. The earliest records of particular groups and the radiations of selected taxa are shown. There is earlier biomarker evidence for eukaryotes around 2,700 Myr (ref. 91). For sources, see text.



bear in mind that initial appearances in the fossil record set only a minimum for the age of clades. Many of the most challenging and controversial questions in evolutionary history concern the origin of major clades (for example, multicellular eukaryotes, animals, land plants, insects and flowering plants), which are also a focus of this review.

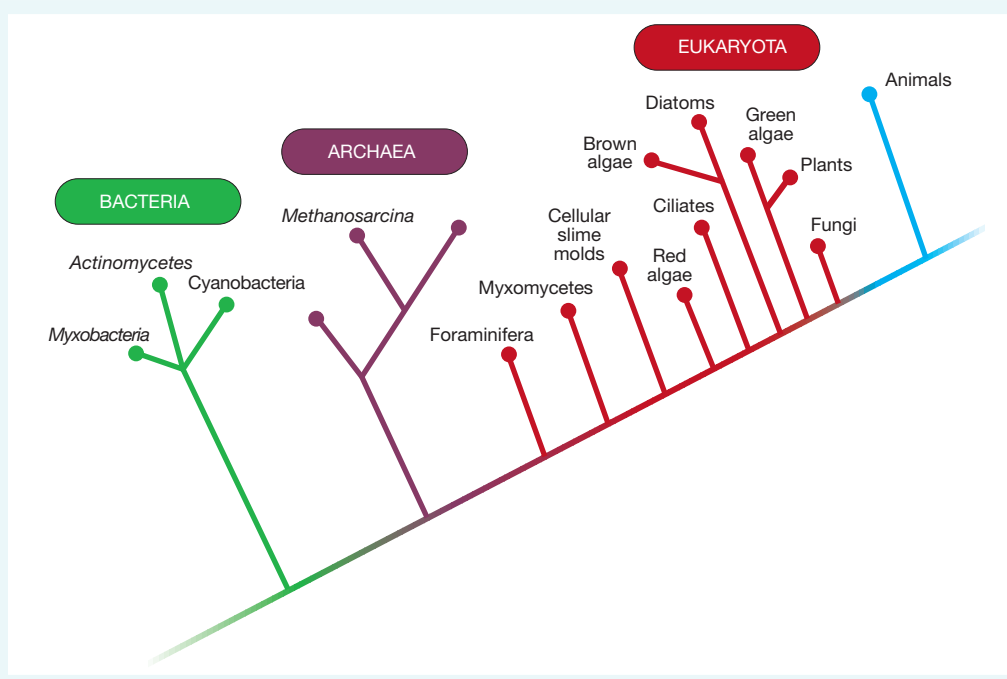
The initial appearances and major radiations of selected taxa documented in the fossil record are summarized in Fig. 1. When considered in the light of the phylogenetic relationships of the major multicellular taxa (Fig. 2), there are three trends evident in the fossil record that I will examine in greater detail. First, multicellularity evolved independently many times and in all three domains of life^{2,13}. Second, following the evolution of multicellularity from different unicellular ancestors, macroscopic forms with new body plans or physiologies and representing higher grades of morphological complexity (for example, multicellular protists, animals and land plants) arose. And third, the emergence of new forms was often followed (after sometimes considerable delays) by periods of rapid diversification (for example, the Cambrian explosion of animals, the rise of insects in the Devonian and Carboniferous, the radiation of flowering plants in the late Cretaceous, and the mammalian radiation in the early Tertiary; Fig. 1). I shall analyse these global trends in the evolution of size, complexity and diversity in more detail and then consider some potential explanations.

Size and multicellularity

For the first 2,500 million years of life on Earth, most species rarely exceeded 1 mm in size and were generally much smaller. The earliest reported bacterial microfossils from about 3,500 million years (Myr) averaged about 5 μm in diameter¹⁴. Early eukaryotic microfossils (acritarchs), while considerably larger, still ranged generally from about 40 to 200 μm in size (with a few larger exceptions, see ref. 15) for much of their first 600–800-Myr history¹⁵. Organismal size increased appreciably with the evolution of multicellular forms. In bacterial and algal forms with cell walls, one of the simplest ways to become multicellular was for the products of cell division to remain together to form long filaments¹³. Many early multicellular eukaryotes were millimetre-scale, linear or branched, filamentous forms^{15,16}.

The size and shape of life did not expand appreciably until the late Proterozoic (Fig. 1). Radially symmetric impressions and trace fossils indicate the presence of millimetre-scale metazoans around 550 Myr (ref. 17; reviewed in ref. 18). The enigmatic Ediacaran fauna comprised of tubular, frond-like, radially symmetric forms generally reached several centimetres in size (although some, such as *Dickinsonia*, approached 1 m), as did macroscopic algae. Organismal sizes expanded considerably in the Cambrian, including bilaterians up to 50 cm in size, as well as sponges and algae up to 5–10 cm (ref. 19). Maximal body lengths of animals increased subsequently

Figure 2 The phylogenetic relationships of multicellular taxa. Multicellularity has evolved independently many times in each domain of life. (Modified from ref. 13.)



by another two orders of magnitude, as did algal sizes (for example, kelp).

The largest extant organisms, giant fungi and trees, evolved from independent small ancestors. Land plants are believed to have evolved from charophyte green algae, and both green algae and plants evolved from a unicellular flagellate ancestor^{20–22}. Fossil spores indicating the earliest evidence of plant life date from the mid-Ordovician. The oldest plant-body fossil (*Cooksonia*) suggests that early land plants were small^{20,21} and, on the basis of molecular phylogenetic analyses, are believed to be comparable in organization and life cycle to liverworts²³. Many of the principal groups of land plants have evolved large (>10 m) species at some point in their history.

Thus, increases in both mean and maximal organismal size occurred in the evolution of multicellular bacteria, eukaryotes and multicellular eukaryotes, and within the algal, fungal, plant and animal lineages.

Complexity

Complexity is one of those problematic terms that has been used to describe so many objects and phenomena as to have lost any generally recognized precision or meaning. In describing organisms, two of the most common usages are in reference to the number of different cell types^{2,13,24} or the number or functional specialization of parts. McShea³ has suggested four distinct categories of complexity that include the specific case of cell-type number but can be used to describe compositions and processes at different levels of biological organization, from molecular to ecological organization. Specifically, these four types of complexity include: (1) the number of different physical parts (for example, genes, cells, organs or organisms) in a system; (2) the number of different interactions among these parts; (3) the number of levels in a causal specification hierarchy; and (4) the number of parts or interactions at a given spatial or temporal scale. These definitions are particularly appropriate in two aspects. First, they allow one to frame questions about the evolution of complexity more narrowly and specifically. Second, they enable the complexity of two independent variables to be compared in order to search for or to refute correlations. For example, it is becoming possible to make comparisons of morphological complexity with the complexity of the genetic systems and developmental programs that generate it.

In each sense of these criteria, the evolution of life has ascended increased grades of complexity. This is most obvious from simple measures of cell number and type (Table 1)^{2,24}. From unicellular

ancestors, multicellular forms have evolved many times in different lineages (Fig. 2). Thus, there have been both global (for example, bacteria to vertebrates) and within-lineage (for example, animals and the green algae/plant clade) increases in the number of cell types. But the maximum number of cell types in general plateaus in bacteria (at 3), in protists (about 4), in protostomes (about 50), and perhaps in vertebrates as well (there is strong suspicion that cell-type number is underestimated in vertebrates but also better studied in these animals).

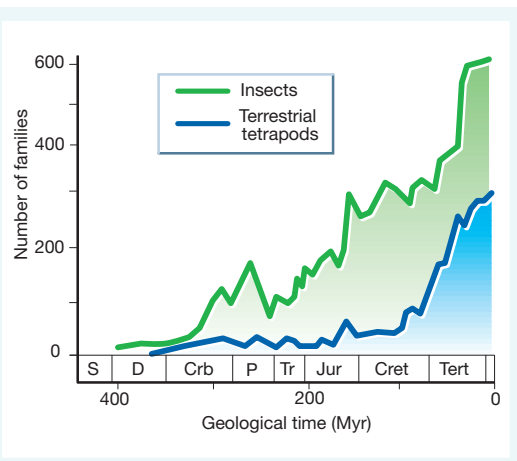
Increases in cell-type number obviously increase the potential physiological and anatomical complexity of organisms through a division of labour among cells and the formation of specialized tissues and organs. Among the bacteria, for example, large cyanobacteria (blue-green algae) have evolved specialized heterocysts that differentiate at regular intervals within a filamentous chain. These heterocysts enable cyanobacteria to segregate the processes of nitrogen fixation (in heterocysts) from photosynthesis (in surrounding vegetative cells)². Spore-forming bacteria are also prokaryotic forms that generate more than one cell type. There are clear advantages to the evolution of hardy spores whose production is under the influence of environmental cues. The shapes of forms with few cell types, from bacteria to slime moulds, are limited generally to filamentous or spherical morphologies. There is little internal morphology in organisms before the evolution of plants, fungi and animals. The evolution of internal complexity accompanied the evolution of greater numbers of cell types and the organization of cells into tissues and organs.

The evolution of cell-type number and internal complexity has been addressed most frequently in the Metazoa, within which the different grades of cellular and anatomical complexity are readily apparent. Placozoans bear only four cell types while the Porifera (sponges) and Cnidaria (including jellyfish and sea anemones) possess 10–12 (ref. 24). Furthermore, cnidarians have only two distinct germ layers (that is, they are ‘diploblastic’), whereas bilaterians possess a third, mesodermal germ layer and considerably more cell types. The evolution of the mesoderm and its derivatives had profound consequences for the evolution of animal body cavities, locomotion and overall size. Among the bilaterians, vertebrates possess the greatest number of cell types, some of which can be attributed to the evolution of the neural crest²⁵.

Another potential index of complexity is gene number. The number of biochemical events within a cell can increase with and therefore bear some correlation to gene number. However, the relationship between gene number and the number of different cell types in multicellular

Figure 3

Patterns of diversification. Terrestrial tetrapods (blue line) and insects (green line) underwent exponential phases of diversification over periods of 100 Myr or more. (Adapted from ref. 30.)



organisms, if any, is not well understood. Recent inventories of the complete genomes of model organisms used for studies of gene regulation, cell differentiation and development offer the opportunity to search for correlations between gene number and cell-type number as indices of complexity.

Among unicellular organisms, total gene numbers range from as few as 470 genes in *Mycoplasma genitalium* to more than 4,000 in *Escherichia coli* (Table 1). There is no apparent relationship between gene number and phylogeny among unicellular organisms. Two members of the Archaea have gene numbers in the middle range of unicellular genomes, as do various bacteria. The smallest genomes are thought to derive from larger genomes through gene loss.

Gene number is not greater in the spore-forming bacterium *Bacillus subtilis* than in other bacteria lacking cell differentiation. However, gene number is considerably greater in the yeast *Saccharomyces cerevisiae* (about 50% greater than the bacterial maximum), which is capable of forming three different cell types (two haploid forms and one diploid form). In the plant *Arabidopsis thaliana* and in two protostomes, the fruitfly *Drosophila melanogaster* and the nematode *Caenorhabditis elegans*, gene number is 2–4 times greater than in *S. cerevisiae*. And in vertebrates, gene number is estimated to be on the order of 4–6 times greater than in the two reported protostome genomes.

From these surveys, we can readily conclude that gene number and cell-type number have increased in the evolution of macroscopic forms from unicellular ancestors. However, the quantitative relationship between the two indices is not at all clear. Gene number varies more than eightfold among unicellular organisms. Although multicellular eukaryotes uniformly possess more genes than bacteria or archaea, organisms with fewer cell types may possess more genes (for example, *A. thaliana* compared with *D. melanogaster*) and the reasons for differences in gene number between species of similar cell-type complexity (for example, *C. elegans* compared with *D. melanogaster*) are unknown. *C. elegans* and *D. melanogaster* belong to the same major protostome clade, the Ecdysozoa, and it is known from analysis of principal developmental genes that the nematode has lost genes that were present in its common ancestors²⁶. In spite of this, the total gene number in *C. elegans* exceeds that of *D. melanogaster* by 5,000 genes. One important contribution to the differences in gene content is the extent and pattern of gene duplications and losses. These events are lineage-specific so the sampling of a few species chosen for reasons other than their phylogenetic relevance may obscure a relationship, if any, between gene number and morphological complexity. However, while total gene number may not be all that informative in regard to complexity, the number of genes with particular developmental functions may be relevant.

Diversity

It is a given that life's diversity has expanded from its origin. The more pertinent question is whether this expansion represents a continuous increase? The answer is most definitely no. Major extinctions have

caused marked reductions in the diversity of the global biota in many episodes of life's history. Furthermore, the dynamics of species diversification and extinction are well studied only for a few groups that have left a long and rich fossil record. For many organisms, particularly those made entirely of soft tissues or of small size, we just cannot say whether total diversity increased or decreased over long periods of time.

What we can say is that there are many episodes in the fossil record of 'bursts' of diversification within lineages. Eukaryotes in the Proterozoic and early Cambrian²⁷, animals in the Cambrian^{28,29}, insects in the Carboniferous, flowering plants in the Tertiary, and other groups experienced periods of rapid radiation (Fig. 1). The geological and ecological settings and the potential catalysts of these periods of accelerated change differ immensely in their particulars.

Perhaps one general theme is that many of these radiations reflect the release from or the surmounting of some environmental or structural constraint(s), or a new way of life. The most obvious of these changes is the transition from an aquatic to a gaseous environment. The invasions of the land by plants and animals were accomplished by enormous changes in physiology and anatomy, which enabled exploitation of new 'ecospace'. The subsequent radiations of land plants, terrestrial tetrapods and insects were explosive and their diversifications followed largely exponential patterns for 100 Myr or longer (Fig. 3)³⁰.

Passive or active global trends?

There are two central questions when evaluating long-term trends in evolution. The first is whether the trend is passive (that is, due to the increase in the total variance in a clade, with the direction of change imposed by the boundary of some initial minimum value), or active (that is, due to the biased replacement of primitive forms with more derived forms). The plot of change in morphology over time distinguishes these two trends (Fig. 4a, b). Given the initial conditions of life (that is, a low minimum boundary), and that simple, small unicellular forms have not been replaced, the global trends in organismal size, complexity and diversity described above must at least in part be passive and due to an overall increase in variance (Fig. 4a). However, the evolution of new traits (such as multicellularity, cell differentiation, internal complexity, support structures and modularity) can establish new levels of complexity and enable subsequent bursts of diversification (through further increases in variance). Thus, whereas global trends may be passive, there may be active, directional trends nested within the overall arc of evolutionary history⁴. To identify these active trends and to consider how they have influenced the shapes of life, we have to look at the histories of individual clades.

The second question in evaluating long-term evolutionary trends, active or passive, concerns whether the mechanism involved is external (affected by selection, ecology or environment) or internal (under genetic, developmental or biomechanical control). It is important to note that the distinction between active and passive trends bears on the question of the potential mechanisms responsible, not on the existence of a trend. Passive trends may well have interesting origins¹⁰. For example, the tendency for mammals to evolve at small sizes ('Cope's rule', see below) begs explanation.

Active trends within clades

Most empirical work on evolutionary trends within clades has focused on animals because of their richer fossil record. Considerable emphasis has been placed upon methodology because different approaches may yield different conclusions regarding the same phenomenon. Three tests have been devised to distinguish passive from active increases: (1) the test of the behaviour of the minimum, which should increase if the system is driven; (2) the ancestor-descendant pair test, which should reveal increases in random samples of ancestor-descendant pairs that are selected away from the boundary of the minimum; and (3) the subclade test, which should reveal a skew in the mean of subclades sampled from the tail of a distribution⁷. These tests have revealed some active trends in the evolution of animal size and complexity.

The evolution of size and Cope's rule

One of the most scrutinized trends is that noted by E. D. Cope³¹ regarding the size increase in mammalian fauna during the Cenozoic (65 Myr–present). Cope attributed the pattern to a tendency for new lineages to evolve at small sizes and an active drive towards increasing size. An active drive has been attributed to various advantages of larger size (for example, evasion of predators, increased brain size or increased longevity; see ref. 8), but rigorous evidence for such a trend was lacking. Stanley⁸ attributed Cope's rule to be due to the tendency for new groups to evolve at small size (that is, an initial minimum value) relative to their optima such that there is a passive drift towards larger mean body size in descendants through an increase in variance. However, Alroy³², in an analysis of body mass estimates for over 1,500 North American fossil mammal species, found a consistent increase between matched pairs of younger and older species. Similarly, MacFadden³³ found a driven trend in the evolution of body size of horses within this era. These within-lineage comparisons indicate that active trends do operate within overall passive trends.

A general applicability of Cope's rule to other taxa and eras has not been found. Neither Cretaceous molluscs³⁴ nor planktonic foraminifera¹⁰ show an active trend. Rather, trends of increasing size in both of these taxa have been attributed to an overall increase in variance.

Complexity of animal morphology

One of the most often analysed trends is the evolution of morphological complexity in the Metazoa. As discussed earlier, cell-type number has increased in the evolution of diploblasts, bilaterians and vertebrates, respectively. As a long-term trend, however, it also seems that cell-type number plateaus in these groups, so there may have been increases in the stem lineages of these groups, but not in the subsequent radiations of these lineages. The discovery of many genetic and developmental similarities in different bilaterian clades has led to the inference of a common ancestor of these clades that was much more anatomically complex than once thought^{18,35,36}. Although likely to have been small in comparison to its Cambrian and later descendants, cell-type number in the last common ancestor of bilaterians was probably comparable to that of modern protostomes and basal deuterostomes³⁷. However, minimum cell numbers probably increased with body size in certain clades in the Cambrian, indicating perhaps a driven trend in the Cambrian radiation.

Complexity with respect to other characters has been scrutinized only for taxa with rich fossil records such as the brachiopods^{3,38}, ammonoids³⁹, aquatic arthropods⁴⁰ and vertebrates⁴¹. Active trends have been identified in the complexity of brachiopod geometry³⁸, ammonoid septal sutures³⁹ and arthropod limb types⁴⁰.

One particular type of complexity of special interest is that of serially repeated structures. Body segments in annelids and arthropods, vertebrate in vertebrates, limbs in many taxa, and teeth are serially homologous structures. Compared to other structures, these are easily quantified and differentiated. Complexity of serially repeated parts is a function then of both the overall total and the number of individual types of structures. In both vertebrates and arthropods there has been a clear increase in the maximum number of distinguishable individual types of repeated structures³. In the evolution of diverse arthropods from trilobitomorphic or lobe-podan⁴² ancestors, the mean and maximum number of distinct limb-pair types increased⁴⁰ (Fig. 5), as did the minimum (temporarily). This suggests a trend that may be driven in part. All three measures have remained static for the past 250 Myr of arthropod evolution (Fig. 5).

In vertebrates as well, the maximum and mean number of differentiated vertebrae has increased at a high taxonomic level in the transition from fish to mammals. Vertebral columns of fossil and modern fish are relatively uniform whereas those of birds and mammals are more complex. Detailed analyses indicate that this is a largely passive trend as the minima have not changed and there is no clear trend of increase in ancestor-descendant pairs⁴¹.

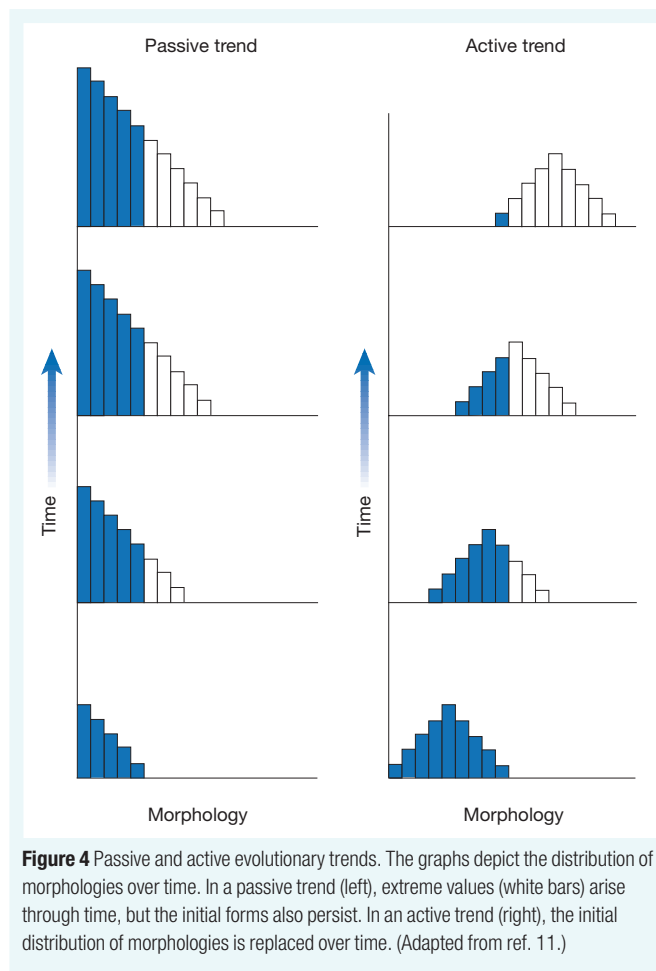


Figure 4 Passive and active evolutionary trends. The graphs depict the distribution of morphologies over time. In a passive trend (left), extreme values (white bars) arise through time, but the initial forms also persist. In an active trend (right), the initial distribution of morphologies is replaced over time. (Adapted from ref. 11.)

Diversity

There are two different elements of diversity that are distinguishable, the morphological and the taxonomic. Morphological diversity is a function of the occupancy of multidimensional morphospace, whereas taxonomic diversity is a function of net speciation events. In principle, these two functions can vary independently of each other. There is empirical evidence that morphological variety and taxonomic diversity can increase together during the initial diversification of a clade⁴³. This trend would be consistent with a passive, diffusive process such that as variance increases, morphospace and ecospace are filled from initial boundary minima.

That the filling of morphospace is initially more rapid is well illustrated by analysis of the evolution of skeletal designs. The 'skeleton space' is a theoretical morphospace against which the actual skeletal designs of fossil and extant organisms have been compared⁴⁴. Of the roughly 180 designs that have been used by all phyla that bear skeletal elements (internal or external, rigid or pliable, uni- or multicomponent, and of various geometries), 146 were exploited by the time of the early Middle Cambrian (in Burgess Shale fauna)⁴⁵. Thus, more than 80% of all designs that ever evolved appeared within the first 6% of overt animal history. Among the earliest and most frequently exploited designs were single-element rods and multielement, metameric exoskeletons. These structures are correlated with the simplest possible strategies for increasing body size. Underrepresented in Cambrian fauna are structures whose growth involves continuous remodelling, such as those of animals with internal skeletons.

Modularity provides the key link in macroform evolution

There is an intuitive relationship between complexity and diversity. As new traits evolve that enable organisms to surmount prior

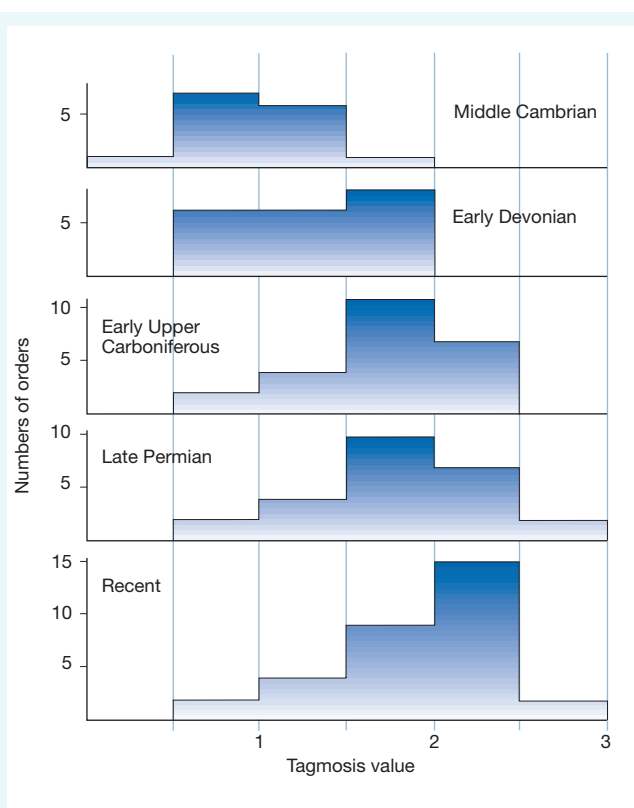


Figure 5 The evolution of limb-type complexity in aquatic arthropods. The number of orders is plotted against tagmosis, a value computed by considering both the total number and different types of arthropod limbs. The minimum, mean, and maximum tagmosis value increased from the Cambrian to the late Permian, suggesting an active trend. (Adapted from ref. 37.)

limitations, such increases in complexity may allow the exploitation of new morphospace and ecospace^{4,46}. In the case of skeletons described above, the evolution of the biochemistry for forming hard parts (chitinous cuticles and mineralized internal or external elements) opened up new ways of life that enabled bursts of diversification. This theme appears repeatedly in evolutionary history. The invasion of the land by plants, tetrapods and insects triggered radiations that transformed terrestrial ecosystems. The early phase of diversification of these groups (considering higher taxonomic levels) was greater than that of later phases, until later innovations (for example, flight in vertebrates or flowers in plants) led to further bursts of diversification. Once again, the overall trends seem to be passive (and those at higher taxonomic levels are to some degree the consequence of how the topology of evolutionary trees are determined^{47,48}), but that does not at all diminish the importance of mechanisms underlying the increases in complexity or diversity in these lineages, the most complex and diverse of all macroscopic forms.

Can we draw any generalities about deeper mechanisms underlying these overall trends, perhaps even the active trends that may be nested within them? I argue here that one of the most critical features underlying the evolution of large and complex animals and plants, and a key to their diversity, is their modular construction. The significance of the construction of animals from repeated parts has long been recognized. Cope³¹, Bateson⁴⁹, Gregory⁵⁰, Rensch⁵¹, Bonner², as well as Darwin (see ref. 52) have suggested various advantages of modular construction, including the facilitation of greater size and efficiency, and the evolution of greater complexity and adaptation through the functional differentiation of repeated parts. Similarly, in plants, modular construction allows for greater size and the differentiation of functional roles among leaf and reproductive structures⁵³. Modularity in plants

and animals can be viewed as being analogous conceptually to the division of labour between cell types of simpler organisms such as cyanobacteria and *B. subtilis*, except that whole body parts are the building blocks rather than cells. The main innovation that enabled large, modular organisms to evolve was the evolution of regional specification systems that subdivide growing embryos into semiautonomous units⁵⁴. We now understand enough about some of the developmental genetic mechanisms for the construction of modular animals (for example, arthropods and vertebrates⁵⁵) and of plant body plans and parts^{56,57} that statements can be made about the mechanistic bases of the morphological diversification of forms at higher taxonomic levels.

For an illustration of the relationship between the evolution of modular body plans, complexity, diversity and the architecture of the underlying genetic systems that differentiate forms, we can return to the example of the evolution of arthropod limb-pair types (Fig. 5). Cisne⁴⁰ documented a trend of an increasing mean and maximum number of limb-pair types in the first half of the Phanerozoic. The functional significance of this potentially driven trend is obvious. Arthropod limbs are “the tools of its trade”⁴⁰ and the morphological specialization of limbs for feeding, locomotion, sensation, copulation, brooding young, burrowing and defence is reflected in the variety of limb types (that is, complexity). The most specialized orders, those with the greatest number of different limb types, are also the most diverse in terms of the number of species. This is illustrated by the malacostracan crustacean lineage that evolved filtering maxilla as a new means of feeding. The evolution of maxillopods freed the serially similar pairs of trunk limbs to become specialized for walking, swimming and burrowing⁴⁰.

The diversification of the modular arthropod body plan and of arthropod limb types required the diversification of the genetic regulatory system that specifies regional and segmental identities in arthropods, namely, the *Hox* genes. The products of the *Hox* genes are expressed in domains that subdivide the anterior–posterior axis of bilaterians, and these proteins regulate the expression of many genes within these domains⁵⁸. The differentiation of serially homologous limbs in arthropods is regulated by different *Hox* genes expressed in different developing limb primordia (reviewed in refs 16, 55 and 59). Comparisons of representatives of all arthropod classes and many different orders have revealed strong correlations between the regional deployment of *Hox* genes and the patterns of limb differentiation (reviewed in refs. 16, 55 and 60). Thus, in taxa with similar sets of repeating limbs (for example, centipedes and brachiopod crustaceans), the same *Hox* genes or combination of *Hox* genes are expressed in all of the limbs of the same type. Whereas in taxa with different numbers and kinds of limbs, relative shifts in *Hox*-gene expression domains are correlated with differences in limb number and identity.

Arthropod limb-type diversity has evolved from ancestors with very similar complements of *Hox* genes⁶¹. This is contrary to initial expectations that gene duplication and divergence would correlate with the duplication and increasing diversity of arthropod segment and limb types^{62,63}. Instead, it is apparent that the great diversification of limb types is due to the evolution of regulatory mechanisms operating at two major hierarchical levels in arthropod development. First, at the level of the regulation of *Hox* genes along the anterior–posterior axis; and second, within the *Hox*-regulated hierarchies of genes that pattern individual limbs. The evolution of axial diversity within the modular body plans of other bilaterian phyla such as annelids and vertebrates has followed similar themes (reviewed in ref. 55).

Modularity, constraints and evolvability

These recent discoveries in developmental genetics and comparative biology illustrate an important property bestowed upon evolving organisms by modularity, that is, the ability to dissociate developmental processes in one part of the body from another^{64,65}. The regional specification systems that subdivide developing animal and plant embryos into discrete territories make it possible for the development and morphology of one territory to evolve independently of

Table 1 Evolution of cell type and gene number

Number of cell-types*	Species	Number of genes in genome	Reference
1	<i>Mycoplasma genitalium</i>	470	76
	<i>Rickettsia prowazekii</i> (intracellular parasite)	834	77
	<i>Haemophilus influenzae</i>	1,709	78
	<i>Escherichia coli</i>	4,288	79
	<i>Campylobacter jejuni</i>	1,654	80
	<i>Aquifex aeolicus</i> (thermophile)	1,512	81
	<i>Neisseria meningitidis</i>	2,121	82
	<i>Archaeoglobus fulgidus</i> (Archaea)	2,436	83
	<i>Methanococcus jannaschii</i> (Archaea)	1,738	84
	<i>Synechocystis</i> sp. (cyanobacterium)	3,168	85
2	<i>Bacillus subtilis</i>	~4,100	86
	<i>Caulobacter crescentus</i>		
3	<i>Saccharomyces cerevisiae</i>	6,241	87
4	<i>Volvox</i>		
	<i>Ulva</i> (sea lettuce) placozoans		
7	Mushrooms		
	Kelp		
~11	Sponge, cnidarians		
~30	<i>Arabidopsis thaliana</i> (plant)	~24,000	88,89
~50	<i>Caenorhabditis elegans</i> (nematode)	18,424	90
	<i>Drosophila melanogaster</i> (fruitfly)	13,601	70
~120	Zebrafish	>80,000–100,000	
~120	Human	~80,000–100,000	

*From refs 2, 24.

another. It has been suggested that modularity (also called compartmentation^{66,67}) facilitates change by conferring upon organisms a greater ability to escape internal constraints on morphology⁶⁷. These constraints include the physical limits imposed by biomechanics on organismal size and shape and genetic and developmental constraints that limit the range of variation that is tolerated and available within species (see ref. 64 for a discussion of constraints). If modularity and dissociation enhance the capacity to generate variation (that is, evolvability), then this may confer a selective advantage on modular clades that possess it^{66,67}. Exploitation of this dissociability is illustrated by the great variety of forms and functions seen in the serially repeated parts of animals and plants.

Genomic complexity and regulatory evolution

The generation of cell-type and body-part diversity depends upon transcriptional regulatory proteins that control the cell- or region-specific expression of target genes. The differences in gene expression between cell types within an organism can number in the hundreds to thousands, but these differences are often controlled by just a small set of regulatory proteins. In *B. subtilis* for example, a small number of regulators control the differential expression of several hundred genes during sporulation⁶⁸. In the yeast *S. cerevisiae*, a small set of transcription factors orchestrates the regulation of genes involved in cell-type differences⁶⁹. In metazoans, cell-type differences (for example, muscle compared with neural) and body-region identity are regulated typically by a few proteins, whereas pattern formation within tissues is regulated by a larger set of proteins⁵⁵. Only about 3–5% of the proteins encoded by animals' genomes are transcriptional regulators^{70,71}. Therefore, the total number of genes is not a driver of cell type or other indices of morphological complexity. For instance, greater gene number (for example, in *C. elegans* compared to *D. melanogaster*, or in zebrafish compared to humans) does not dictate greater cell-type number or any other index of morphological complexity (indeed, *C. elegans* is a highly derived ecdysozoan that has lost certain morphological features (eyes) and *Hox* genes that were present in the common ancestor it shares with *Drosophila*).

It seems that most expansion in the genetic toolkit for bilaterian development occurred in two intervals of bilaterian evolution. First, in the stem lineage leading to bilaterians, and later, in the evolution of vertebrates from a chordate ancestor⁷². These two intervals of *Hox* gene expansion do correlate with the evolution of increased grades of

complexity. Interestingly, protostomes and more basal deuterostomes possess similar complements of regulators that control cell-type, tissue and regional identities.

Expansions in the number of regulatory proteins offer the potential, but are not necessary for, the evolution of increased complexity and the expansion of diversity. The evolutionary trends in arthropod and vertebrate axial complexity and diversity are due to the evolution of genomic complexity at a different level than gene number, that is, at the level of the evolution of the regulatory elements that act in *cis* to control gene expression. Within these phyla, no substantial increase in the number of genes involved in regional specification occurred. Rather, the evolution of advanced forms with greater numbers of differentiated serial structures has occurred through an expansion in the number of regulatory elements that control region-specific expression of genes. The expansion of regulatory elements constitutes increases in genomic complexity in all four senses described earlier — in the number of different parts (regulatory elements) in a regulatory system, in the number of different interactions of these parts, in the number of levels in developmental hierarchies, and in the number of parts and interactions at a given spatial scale. Regulatory evolution creates new combinations of gene expression and therefore enables increases in the information content of genomes and the generative potential of development without expansion of gene number. The role of regulatory evolution is therefore key to understanding how morphological complexity and diversity evolve in macroscopic forms.

Chance and necessity

Perhaps the most surprising conclusion one might draw from the consideration of the complexity of genes and genomes is that the generative potential of genomes is far greater than is realized in evolution. Kauffman⁷³ has pointed out that there is a vast difference between the potential number of combinations of possible gene expression states and those that actually exist in any organism. Given just two inputs into each gene, a system of 100,000 genes has $2^{100,000}$ different possible states. Yet, if we use cell-type number as an indicator of gene expression states, only 200–300 states are realized (more states are realized in modular organisms through differential expression in different modules). Furthermore, given that multicellularity seems to be readily evolved and that very few regulatory proteins can orchestrate markedly different cell physiologies, it is curious that more multicellular forms have not evolved.

We do not understand why the actual complexity realized in evolution is far less than what seems to be possible genetically. The observed limits of form seem to be due to a combination of both chance and necessity, a product of historical contingency and imposed by external agents (for example, selection) and internal rules (for example, constraints). The demands of natural selection may exclude or favour certain forms, but it is widely agreed that selection cannot be the whole story. Internally imposed constraints also shape the world of possible morphologies and are themselves factors that can evolve⁷⁴.

Are there universal rules to the shapes of life?

The parochial question nested within the mystery of the existence of life on other bodies is that of the existence of forms like the ones that have occurred on Earth. A few extrapolations seem to be reasonably grounded in the overall trends of life's history reviewed here. Assuming a cellular basis of life elsewhere, the passive trends towards increases in organismal size, complexity and diversity from some initial minima are certain to prevail in any system. It must be kept in mind, however, that few macroscopic forms evolved in the first 3 billion years of life on Earth. Therefore, the time required for any quantum change in morphology is entirely contingent upon the particular history of any system. As for the shapes of life, macroscopic forms are most likely to be multicellular and there is a finite set of simple geometries — such as those that dominated the early history of life on Earth (linear and branched filaments, cylinders and

spheres) — that are likely to satisfy the constraints imposed by diffusion and biomechanics and that are therefore likely to be universal^{2,75}.

But the evolution of motile, modular mega-organisms may be a different story. Only after 3 billion years of physiological and anatomical evolution, vast changes in the environment and ecology (that were partly biogenic in nature), and extensive genetic and developmental innovations did such beasts emerge on Earth. And, although some symmetrical body organization is likely of macro-forms⁷⁵, there is no basis to assert that bilateral, radial or spiral forms were or would be inevitable. Nor, sadly, is their continued evolution assured as the ecological dice are now in the hands of a single species that is on a path to extinguishing a substantial fraction of all diversity before the question of life elsewhere may be answered. □

1. Monod, J. *Chance and Necessity* (Vintage Books, New York, 1971).
2. Bonner, J. T. *The Evolution of Complexity by Means of Natural Selection* (Princeton Univ. Press, Princeton, 1988).
3. McShea, D. W. Metazoan complexity and evolution: is there a trend? *Evolution* **50**, 477–492 (1996).
4. Knoll, A. H. & Bambach, R. K. Directionality in the history of life: diffusion from the left wall or repeated scaling of the right? *Paleobiology* **26** (Suppl.) 1–14 (2000).
5. Gould, S. J. *Full House* (Harmony Books, New York, 1996).
6. Foote, M. Cambrian and recent morphological disparity. *Science* **258**, 1816–1818 (1992).
7. McShea, D. W. Mechanisms of large-scale evolutionary trends. *Evolution* **48**, 1747–1763 (1994).
8. Stanley, S. M. An explanation for Cope's rule. *Evolution* **27**, 1–26 (1973).
9. Maynard Smith, J. Time in evolutionary process. *Studia Generale* **23**, 266–272 (1970).
10. Gould, S. J. Presidential Address. Trends as changes in variance: a new slant on progress and directionality in evolution. *J. Paleont.* **62**, 319–329 (1988).
11. Wagner, P. J. Contrasting the underlying patterns of active trends in morphological evolution. *Evolution* **50**, 990–1007 (1996).
12. Hawking, S. W. *A Brief History of Time: From the Big Bang to Black Holes* (Bantam Books, Toronto, 1988).
13. Bonner, J. T. The origins of multicellularity. *Integ. Biol.* **1**, 28–36 (1998).
14. Schopf, J. W. Microfossils of the Early Archean Apex chert: new evidence of the antiquity of life. *Science* **260**, 640–645 (1993).
15. Knoll, A. H. The early evolution of eukaryotes: a geological perspective. *Science* **256**, 622–627 (1992).
16. Butterfield, N. J. *Bangiomorpha pubescens* n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. *Paleobiology* **26**, 386–404 (2000).
17. Martin, M. W. *et al.* Age of Neoproterozoic bilaterian body of trace fossils, White Sea, Russia: implications for metazoan evolution. *Science* **288**, 841–845 (2000).
18. Knoll, A. H. & Carroll, S. B. Early animal evolution: emerging views from comparative biology and geology. *Science* **284**, 2129–2137 (1999).
19. Briggs, D. E. G., Erwin, D. H. & Collier, F. J. *The Fossils of the Burgess Shale* (Smithsonian Institution Press, Washington DC, 1994).
20. Kenrick, P. & Crane, P. R. *The Origin and Early Diversification of Land Plants* (Smithsonian Institution Press, Washington DC, 1997).
21. Graham, L. E. *Origin of Land Plants* (Wiley, New York, 1993).
22. Kirk, D. L. *Volvox: Molecular Genetic Origins of Multicellularity and Cellular Differentiation* (Cambridge Univ. Press, Cambridge, 1998).
23. Qui, Y.-L., Cho, Y., Cox, J. C. & Palmer, J. D. The gain of three mitochondrial introns identifies liverworts as the earliest land plants. *Nature* **394**, 671–674 (1998).
24. Valentine, J. W., Collins, A. G. & Meyer, C. P. Morphological complexity increase in metazoans. *Paleobiology* **20**, 131–142 (1994).
25. Shimeld, S. M. & Holland, P. W. H. Vertebrate innovations. *Proc. Natl Acad. Sci. USA* **97**, 4434–4437 (2000).
26. de Rosa, R. *et al.* Hox genes in brachiopods and priapulids and protostome evolution. *Nature* **399**, 772–776 (1999).
27. Knoll, A. H. Proterozoic and early Cambrian protists: evidence for accelerating evolutionary tempo. *Proc. Natl Acad. Sci. USA* **91**, 6743–6750 (1994).
28. Conway Morris, S. *The Crucible of Creation: The Burgess Shale and the Rise of Animals* (Oxford Univ. Press, Oxford, 1998).
29. Budd, G. E. & Jensen, S. A critical reappraisal of the fossil record of the bilaterian phyla. *Biol. Rev.* **75**, 253–295 (2000).
30. Benton, M. J. Models for the diversification of life. *Trends Ecol. Evol.* **12**, 490–495.
31. Cope, E. D. The method of creation of organic forms. *Proc. Am. Phil. Soc.* **12**, 229–263 (1871).
32. Alroy, J. Cope's rule and the dynamics of body mass evolution in North American fossil mammals. *Science* **280**, 731–734 (1998).
33. MacFadden, B. J. Fossil horses from "Eohippus" (Hyracotherium) to Equus: scaling, Cope's law, and the evolution of body size. *Paleobiology* **12**, 355–369 (1986).
34. Jablonski, D. Body-size evolution in Cretaceous molluscs and the status of Cope's rule. *Nature* **385**, 250–252 (1997).
35. De Robertis, E. M. & Sasaki, Y. A common plan for dorsoventral patterning in Bilateria. *Nature* **380**, 37–40 (1996).
36. De Robertis, E. M. The ancestry of segmentation. *Nature* **387**, 25–26 (1997).
37. Valentine, J. Late Precambrian bilaterians: grades and clades. *Proc. Natl Acad. Sci. USA* **91**, 6751–6757 (1994).
38. Carlson, S. J. Evolutionary trends in the articulate brachiopod hinge mechanism. *Paleobiology* **18**, 344–366 (1992).
39. Saunders, W. B., Work, D. M. & Nikolaeva, S. V. Evolution of complexity in Paleozoic ammonoid structures. *Science* **286**, 760–763 (1999).
40. Cisne, J. L. Evolution of the world fauna of aquatic free-living arthropods. *Evolution* **28**, 337–366 (1974).
41. McShea, D. W. Evolutionary change in the morphological complexity of the mammalian vertebral column. *Evolution* **47**, 730–740 (1993).
42. Budd, G. The morphology of *Opabinia regalis* and the reconstruction of the arthropod stem-group. *Lethaia* **29**, 1–14 (1996).
43. Foote, M. Discordance and concordance between morphological and taxonomic diversity. *Paleobiology* **19**, 185–204 (1993).
44. Thomas, R. D. K. The skeleton space: a finite set of organic designs. *Evolution* **47**, 341–360 (1993).
45. Thomas, R. D. K., Shearman, R. M. & Stewart, G. W. Evolutionary exploitation of design options by the first animals with hard skeletons. *Science* **288**, 1239–1242 (2000).
46. Stebbins, G. L. *The Basis of Progressive Evolution* (Univ. North Carolina Press, Chapel Hill, 1969).
47. Raup, D. M. On the early origins of major biologic groups. *Paleobiology* **9**, 107–115 (1983).
48. Smith, A. B. *Systematics and the Fossil Record* (Blackwell Scientific Publications, Oxford, 1994).
49. Bateson, W. *Materials for the Study of Variation* (Macmillan, London, 1894).
50. Gregory, W. K. Reduplication in evolution. *Quart. Rev. Biol.* **10**, 272–290 (1935).
51. Rensch, B. *Evolution above the Species Level* (Columbia Univ. Press, New York, 1960).
52. McShea, D. W. Complexity and evolution: what everybody knows. *Biol. Philosophy* **6**, 303–324 (1991).
53. Niklas, K. J. *Plant Allometry: The Scaling of Form and Process* (Univ. Chicago Press, Chicago, 1994).
54. Davidson, E., Peterson, K. & Cameron, R. Origin of bilaterian body plans: evolution of developmental regulatory mechanisms. *Science* **270**, 1319–1325 (1995).
55. Carroll, S. B., Grenier, J. & Weatherbee, S. D. *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design* (Blackwell, Cambridge, 2001).
56. Ma, H. & DePamphilis, C. The ABCs of floral evolution. *Cell* **101**, 5–8 (2000).
57. Ambrose, B. A. *et al.* Molecular and genetic analyses of the *silky1* gene reveal conservation in floral organ specification between eudicots and monocots. *Mol. Cell* **5**, 569–579 (2000).
58. Carroll, S. Homeotic genes and the evolution of arthropods and chordates. *Nature* **376**, 479–485 (1995).
59. Weatherbee, S. D. & Carroll, S. B. Selector genes and limb identity in arthropods and vertebrates. *Cell* **97**, 283–286 (1999).
60. Akam, M. Arthropods: developmental diversity within a (super) phylum. *Proc. Natl Acad. Sci. USA* **97**, 4438–4441 (2000).
61. Grenier, J., Garber, T., Warren, R., Whittington, P. & Carroll, S. Evolution of the entire arthropod *Hox* gene set predated the origin and radiation of the onychophoran/arthropod clade. *Curr. Biol.* **7**, 547–553 (1997).
62. Lewis, E. B. A gene complex controlling segmentation in *Drosophila*. *Nature* **276**, 565–570 (1978).
63. Akam, M. The molecular basis for metameric pattern in the *Drosophila* embryo. *Development* **101**, 1–22 (1987).
64. Raff, R. A. *The Shape of Life* (Univ. Chicago Press, Chicago, 1996).
65. Raff, R. & Sly, B. J. Modularity and dissociation in the evolution of gene expression territories in development. *Evol. Dev.* **2**, 102–113 (2000).
66. Gerhart, J. & Kirschner, M. *Cells, Embryos, and Evolution* (Blackwell Science, Malden, 1997).
67. Kirschner, M. & Gerhart, J. Evolvability. *Proc. Natl Acad. Sci. USA* **95**, 8420–8427 (1998).
68. Fawcett, P., Eichenberger, P., Losick, R. & Youngman, P. The transcriptional profile of early to middle sporulation in *Bacillus subtilis*. *Proc. Natl Acad. Sci. USA* **97**, 8063–8068 (2000).
69. Johnson, A. D. Molecular mechanisms of cell-type determination in yeast. *Curr. Opin. Genet. Dev.* **5**, 552–558 (1998).
70. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
71. Ruvkin, G. & Hobart, O. The taxonomy of developmental control in *C. elegans*. *Science* **282**, 2033–2041 (1998).
72. Garcia-Fernandez, J. & Holland, P. W. Archetypal organization of the amphioxus *Hox* gene cluster. *Nature* **370**, 563–566 (1994).
73. Kauffman, S. *At Home in the Universe* (Oxford University Press, New York, 1995).
74. Budd, G. E. Does evolution in body patterning genes drive morphological change—or vice versa? *BioEssays* **21**, 326–332 (1999).
75. Stewart, I. *Life's Other Secret* (Wiley, New York, 1998).
76. Fraser, C. M. *et al.* The minimal gene complement of *Mycoplasma genome*. *Science* **270**, 397–404 (1995).
77. Andersson, S. G. E. *et al.* The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133–143 (1998).
78. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–412 (1995).
79. Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462 (1997).
80. Parkhill, J. *et al.* The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403**, 665–668 (2000).
81. Deckert, G. *et al.* The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392**, 353–358 (1998).
82. Parkhill, J. *et al.* Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* **404**, 502–506 (2000).
83. Klenk, H.-P. *et al.* The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390**, 364–370 (1997).
84. Bult, C. J. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058–1073 (1996).
85. Kaneko, T. *et al.* Sequence analysis of the genome of the unicellular Cyanobacterium *Synechocystis* sp. strain. *DNA Res.* **3**, 109–136 (1996).
86. Kunst, F. *et al.* The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256 (1997).
87. Goffeau, A. *et al.* The yeast genome directory. *Nature* **387**, 5–105 (1997).
88. European Union Arabidopsis Genome Sequencing Consortium & Cold Spring Harbor, Washington University in St Louis and PE Biosystems Arabidopsis Sequencing Consortium. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* **402**, 769–777 (1999).
89. Lin, X. *et al.* Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402**, 761–768 (1999).
90. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215 (2000).
91. Brocks, J. J., Logan, G. A., Buick, R. & Summons, R. E. Archean molecular fossils and the early rise of eukaryotes. *Science* **285**, 1033–1036 (1999).

Acknowledgements

I thank A. Knoll for discussions and helpful pointers; R. Losick and A. Johnson for information on microbial gene regulation; G. Budd, J. Crow, N. King and J. True for suggestions on the text; J. Carroll for preparation of the manuscript; and L. Olds for the artwork. S.B.C. is an Investigator of the Howard Hughes Medical Institute.

The search for extraterrestrial intelligence

T. L. Wilson

Max-Planck-Institut für Radioastronomie, Auf dem Huegel 69, D 53121 Bonn, Germany, and Sub-Millimeter Telescope Observatory, Steward Observatory, The University of Arizona, Tucson, Arizona, USA (e-mail twilson@as.arizona.edu)

As far as we know, humanity is alone in the Universe: there is no definite evidence for the existence of extraterrestrial life, let alone extraterrestrial civilizations (ETCs) capable of communicating or travelling over interstellar distances. Yet popular speculation about the existence of ETCs abounds, including reports of alien visitations either now or in the past. But there is a middle way. It is now possible to put limits on the existence of ETCs of varying capabilities, within arbitrary distances from the Solar System, and conceive of real-world strategies whereby we might communicate with ETCs, or they with us.

One of the intriguing aspects of astrobiology in all its forms is that matters once confined to science fiction are increasingly being discussed seriously by scientists. SETI — the Search for Extraterrestrial Intelligence — is perhaps the best example. Knowledge of the space environment gathered over the past 40 years, as well as improvements in our own technological capabilities, can now be applied to make plausible estimates of the number of ETCs in the Galaxy with technologies comparable to or more advanced than ours. Even so, these estimates vary greatly^{1,2}. Recent developments and discussion could help constrain such estimates, including the discovery and characterization of planets around other stars³; reports of the discovery of possible past life on Mars⁴; and Ward and Brownlee's book *Rare Earth*⁵, in which they argue that life is common in the Universe, but ETCs are rare.

A history of SETI

When modern considerations of SETI began, microwave radar and radio astronomy in the centimetre wavelength range were mature technologies. Thus it was natural to believe that an ETC might transmit signals with radio telescopes similar to those then in operation^{6,7}. SETI began in 1960 with the targeted search of two nearby Sun-like stars, using the 25-metre-diameter radio telescope of the National Radio Astronomy Observatory⁷. This was Project OZMA, named after the queen in an imaginary land, "very far away and populated by strange and exotic beings"⁸. OZMA was soon followed by other initiatives, all using radio telescopes⁹, and SETI programmes continue today, despite uncertain funding. NASA's programme, the High Resolution Microwave Search (HRMS) included both a targeted search (examination of selected stars) and an all-sky survey, but was cancelled by the US Congress in the early 1990s¹⁰. Funding for SETI initiatives today comes from non-profit organizations such as the Planetary Society¹¹ and the SETI Institute¹². The SETI Institute has revived the targeted-search portion of the HRMS as 'Project Phoenix' using systems based on the NASA detectors. Meanwhile, SERENDIP (Search for Extraterrestrial Radio Emissions from Nearby Developed Intelligent Populations) — a programme at the University of California, Berkeley — provides detectors for large radio telescopes so that ETC searches can be made while conventional astronomical investigations are carried out¹³.

SETI that does not involve direct physical contact with ETCs (which I rule out for the purposes of this article) is based usually around three rather different scenarios. First, the detection of electromagnetic signals from an ETC, deliberately targeted at us with the expressed intention of communication; second, the detection of signals from an ETC targeted elsewhere, but still designed for communication (for example, a space navigational beacon); and third, the detection of stray electromagnetic radiation from an ETC, not intended for interstellar communication (for example, radio or television signals). These scenarios increase in plausibility as well as difficulty, and are not mutually exclusive, given the difficulty of judging the intentions of ETCs.

A SETI practical

Before we even begin a search, we must be clear about the physical limitations of the technology available to us, both for the reception of signals and their transmission by ETCs. But within these limitations lie certain strengths, for they allow us to set limits on the existence of ETCs capable of specific technological capabilities within arbitrary distances from the Earth.

Early SETI projects concentrated on listening for electromagnetic signals in the centimetre waveband, between around 3 and 60 cm. The reasons for this choice are purely practical: it is in this region of the radio spectrum that background noise from the Galaxy, the Earth's atmosphere and the receiving equipment is lowest. Figure 1 shows the contributions from various sources of naturally occurring noise. Some of the basic concepts of the single parabolic reflectors used conventionally in radio astronomy¹⁴, and which could be used to transmit as well as receive signals, are summarized in Fig. 2. When used for receiving transmissions, large antennas have the advantage of greater collecting area, and parabolic reflectors reap the rewards of greater sensitivity when targeted to specific directions. For example, a parabolic reflector of diameter 100 m can detect signals that are ten-millionths the strength of signals detectable to a non-directional antenna.

The same principle applies to the transmission of signals. Antennas directed at small regions are much more efficient (they have greater effective radiated power, or ERP) than antennas radiating isotropically, and transmission is better at shorter wavelengths. For example, targeted transmission at a wavelength of around 1 cm is 100 times as effective as transmission at 10 cm: power is more concentrated at shorter wavelengths, allowing more effective propagation of a signal

through space. However, the optimal waveband for signal transmission does not coincide exactly with the 3–60-cm band used in early SETI projects, and it is easy to see why. The background noise generated by a receiver operating at 1 cm is tenfold that at 10 cm, and the Earth's atmosphere is more of an obstacle. The atmosphere becomes a serious problem at wavelengths shorter than 1 cm, but this could be circumvented by siting radio telescopes at high altitude or in space.

Which wavelength?

Since the days of Project OZMA, receiver sensitivity at centimetre wavelengths has improved 20-fold. In OZMA, 100-Hz 'windows' in the spectrum were searched sequentially⁷. Now, large wavelength regions are analysed simultaneously by parallel sets of detectors. This allows increases in the speed of searches by a factor equal to the number of channels, and soon it will be possible to analyse a billion contiguous, narrow channels simultaneously¹⁵. Expansion is possible in another direction — using an array of coupled radio antennas, rather than a single dish, to create a radio interferometer¹⁴. Such an array would be cheaper to build than a single telescope of the same size, and could also be used to receive signals from several different regions simultaneously. The Allen Telescope Array (ATA), funded by private sources and to be operated by the SETI Institute, is one such instrument. Both SETI and conventional radio astronomy projects could be carried out with this array, which will operate in the 3–60-cm waveband¹⁶.

But the SETI searchers of the 1960s had one big advantage over contemporary scientists — they did not have to worry about the waveband they chose to examine. The 3–60-cm band was convenient, and within that band the choice was simple: in 1960, only one spectral line was known to be present in this region, namely the well known 21-cm emission line of neutral hydrogen. The hydrogen line is a prominent and universal feature, and proponents of SETI hope that ETCs would emit signals using carriers at or near this wavelength, taking advantage incidentally of the fact that it falls within the relatively radio-quiet 3–60-cm range. Since then, it has been found that the hydroxyl radical (OH) emits at 18 cm, so the 18–21-cm region — associated with the components of water — has been dubbed the 'water hole', a feature that might be exploited by ETCs seeking to communicate with life forms for whom water would be important¹⁷.

Modern SETI is spoiled for choice: tens of thousands of spectral lines

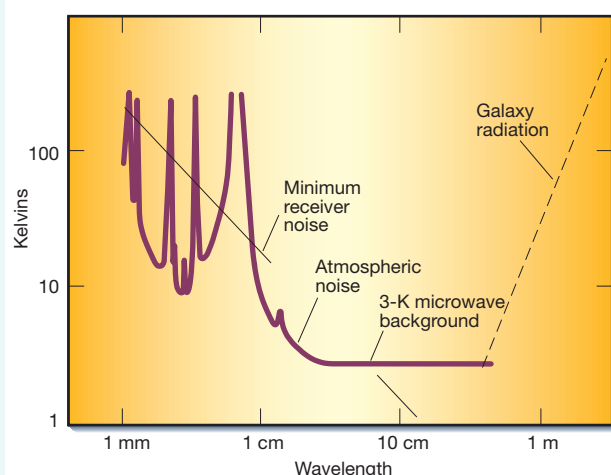


Figure 1 Noise from the Earth's atmosphere and the minimum noise from our Galaxy, versus wavelength. The vertical axis is in kelvins, which is proportional to power per unit bandwidth. The plot illustrates the low noise floor of the region between 3–60 cm, explaining its popularity among proponents of SETI. The curve marked 'minimum receiver noise' is ten times the theoretical minimum for radio-type receivers; the noise of many radio-type receivers approaches this value. The 3-K microwave background is the remnant radiation from the Big Bang¹⁴. Interstellar dust absorbs optical radiation, but not radio waves, so in the plane of our Galaxy radio signals can penetrate to great distances¹⁴.

Box 1

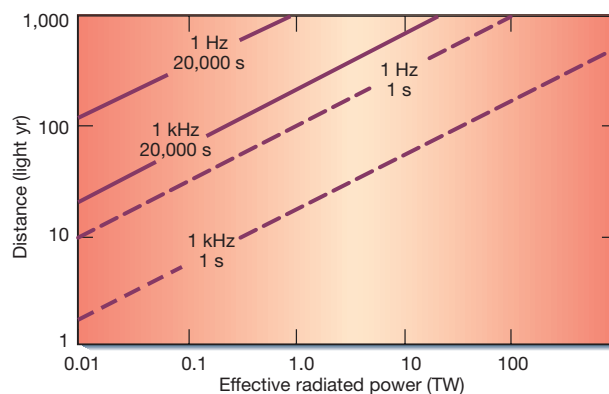
Power requirements

The measurement of interest for a transmitter–antenna combination is the effective radiated power (ERP). ERP is proportional to the product of the effective power fed to the antenna and the squared diameter of the antenna in wavelengths. This can amount to a very large amount of energy: for example, if a 1-MW transmitter operating at 10 cm is connected to a 100-m antenna, the peak ERP is equivalent to the power used for all technological activities on Earth around 1970. And if there were another 100-m radio telescope receiving this signal at a distance of 100 light years, detection would occur in less than 30 s.

Spreading the bandwidth makes detection more difficult, as the figure below shows. A signal transmitted over a 1-Hz band can be detected from much further away than a signal having the same ERP but transmitted in a bandwidth of 1 kHz. If an ETC is sending a constant signal, then we increase our sensitivity by summing the result in time. Such a process increases the distance at which we can detect a given signal with a given ERP. Summing the signal for 20,000 s instead of 1 s allows us to detect a transmitter 12 times as distant.

All the above assumes that signals from ETCs would be sent intentionally for the purposes of communication. Much more problematic is the detection of 'inadvertent' signals from an ETC, such as from television transmitters³³, interstellar navigation beacons or the microwave emissions from orbiting satellites used for solar power production³⁴.

As an example, at a distance of just 10 light years, and using a 100-m telescope, we would need to sum a 6-MW signal for 10,000 years to detect it. We can shorten this time by increasing the size of the telescope. There was an ambitious proposal to build 'Project Cyclops', an array of up to 1,500 radio telescopes each of 100-m diameter¹⁷. With Cyclops, television broadcasts from ETCs could be detected by summing the signal for 4 h. But even in this optimistic case, detection of an inadvertent signal would not be easy.



Box 1 Figure Signal detection as a function of transmitter power, signal bandwidth and integration time. The diagonal lines represent the distance at which a given signal can be detected at five times the noise for a given bandwidth (in hertz) and a given integration time (in seconds). If, for example, a 200-MW transmitter operating at 3 cm is coupled to a 100-m radio telescope with an efficiency of 0.6, the ERP is 10^4 TW (10^{10} MW). The calculation is based on the assumption that the receiving antenna has a diameter of 100 m, an efficiency of 0.6, and is equipped with a 20-K noise temperature receiver. There are two sets of curves, for bandwidths of 1 Hz and 1 kHz. The dashed curves shown have shorter integration times than the corresponding solid curves¹⁸.

Box 2

The Kardashev classification

Kardashev²¹ classified possible ETCs according to the energy at their disposal (see table below). This scheme allows us to determine whether we are dealing with a civilization like our own (type I), a rather advanced civilization (type II) or a vastly more advanced civilization (type III). Humanity has sufficient resources at present to broadcast messages comparable to a type I civilization in a specific direction, although in practice the types of transmission are based on isotropic radiators. A Type II transmission might be transmitted by an ETC that had captured all of the power from its central star. These ETCs are referred to as Dyson civilizations²⁷. Type III civilizations have captured the power of an entire galaxy.

In a recent survey of SETI results, it is reported that searches rule out type II civilizations to a distance of 10 million light years and type I civilizations to 1,000 light years (ref. 22). However, this is based on two assumptions. First, that ETCs transmit at centimetre-scale radio wavelengths, and second, that the bands surveyed include the transmitter wavelengths of nearby galaxies. Conservatively, one can state that for a sizeable part of our galaxy we can probably rule out the presence of type II and III civilizations, if these ETCs are broadcasting messages in the centimetre wavelength range. However, there exist sources of radio emission that are about a million times as powerful as those at the disposal of hypothetical type I ETCs. These arise from regions smaller than 1,000 times the distance between the Earth and Sun, although a detailed analysis of the noise characteristics of these signals shows that they are natural sources of radiation³⁵.

Box 2 Table Transmission power of ETCs

Civilization	ETC transmission power
Type I	The power expended by all technological activity on Earth. For a specific direction, this can be achieved by coupling the output of a 1-MW transmitter, operating at 10 cm, to a 100-m diameter telescope.
Type II	The power in the entire output of the Sun. This is equal to 10^{14} times a type I transmission.
Type III	The power from our entire Galaxy. This is 10^{11} times a type II signal.

from more than 115 known interstellar molecules are now known, so that the choice of a SETI wavelength has become complex. Many authors argue that the 18–21-cm ‘water hole’ range is best, but this in itself does not rule out other possibilities⁷. The intentions of ETCs are always a wild card: ETCs might choose to send messages at a wavelength obtained from various combinations of atomic constants. One such combination gives a wavelength of 11.72 cm (ref. 18), but there is no compelling argument for this wavelength either. Because agreement is lacking on which wavelength is the most likely, one approach is a systematic search through the range from 3 to 60 cm. An added advantage of using this range is that it is easier to build radio telescopes with large collecting areas. But as already noted, there are also arguments for specific frequencies outside this range. At shorter wavelengths, other choices would be 1.35 cm (a transition of water) or 1.47 mm (the analogue of the 21-cm line for a positron–electron ‘atom’^{19,20}).

Effective power

Successful transmission and reception are limited by available power (Box 1). If the distance at which a signal can be detected by a 100-m radio telescope is plotted as a function of transmitter power, signal bandwidth and integration time (Box 1 Figure), it is clear that the most effective radio transmission must use a narrow bandwidth. The drawback, familiar to any Internet user, is that the narrower the bandwidth, the slower the transmission of information¹⁴. Of course, we cannot know *a priori* the power at the disposal of any proposed ETC. As a way of addressing this, Kardashev²¹ devised a scheme for the classification of ETCs according to the resources they might be able to command. A

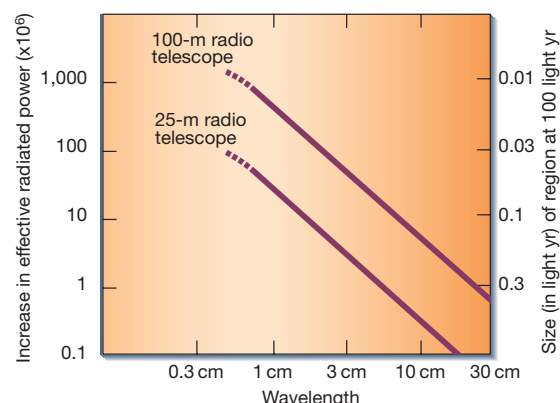


Figure 2 How increases in effective radiated power (ERP) of a transmitting radio dish are related to wavelength and the size and distance of the target. The increase in ERP is plotted on the vertical axis for two antennas as a function of wavelength. For a given antenna size, the power is increasingly concentrated into a smaller region as the wavelength decreases. When the antenna surface becomes comparable to the wavelength, this dependence breaks down (indicated by the dashed line). The 100-m diameter antenna has 16 times the effective power of a 25-m antenna for the same wavelength and target size. This advantage also holds when the telescope is used for receiving: the size of the region contained within the telescope beam at a distance of 100 light years is shown on the right-hand scale. For example, the 100-m radio telescope will include a region 0.14 light years across. Because this is about 4,000 times the diameter of the Earth’s orbit around the Sun, we would certainly receive a transmission from an Earth-like planet when pointing towards a star at 100 light years.

‘type I’ civilization controls the resources of its home planet. Humanity is close to that point now. A ‘type II’ civilization can control the output of its home star, whereas a ‘type III’ civilization can dispose of the resources offered by its home galaxy (see Box 2). This classification can be used to set bounds on the existence of ETCs of specified technological capabilities within arbitrary distances of the Earth.

Given that we seek to receive and/or transmit signals from a specific direction, rather than isotropically, it should be possible in principle to exchange signals with a type I ETC using a 100-m radio telescope, together with a high-power transmitter to send narrow-band signals. These would be detectable with another 100-m radio telescope even out to a distance of 1,000 light years. A message from a type II civilization, on the other hand, would be detectable with a telescope of 10-m diameter from a distance of 100 million light years, which is 100,000 times further than the limit for a type I ETC. Type III ETCs, with galaxies of energy at their command, should be detectable to distances 10,000 times greater than the most distant type II ETC.

Stars are concentrated in galaxies, and there are more than 20 galaxies within 3 million light years of the Milky Way. In principle, we should be able to receive a message from type II or III ETCs in any of these with technology currently available. With an average of 10,000 million Sun-type stars per galaxy, we could detect messages from ETCs even if the product of the last five terms in the Drake equation, which computes the number of communicating civilizations in a galaxy at a given time (see Box 3), were less than one part in a 100 million. These considerations provide a rationale for all-sky, untargeted searches: with the possibility of at least modest numbers of perhaps readily detectable ETCs (especially of type II or III), the extra sensitivity conferred by targeted searches would not be an absolute requirement for success. However, the fact remains that no confirmed transmissions in the centimetre wavelength range have been received^{19,22}, from which it has been claimed that type II and type III ETCs do not exist at the present epoch²².

This claim is overstated: it may be valid for a sizeable part of our Galaxy, but only if the ETCs are broadcasting in the centimetre wavelength range without interruption — and if they wish their signals to be detectable. Signals in the centimetre range might be attenuated or

Box 3

The Drake equation

The Drake equation is an attempt to quantify estimates of the number of ETCs⁷. This relation is

$$N = R \cdot f_p \cdot n_e \cdot f_i \cdot f_c \cdot L$$

where N is the number of ETCs communicating at any given time; R is the average rate of galactic star formation; f_p is the fraction of stars accompanied by planets; n_e is the number of planets per star system with conditions needed to support life; f_i is the fraction of habitable planets on which life actually arises; f_c is the fraction of the life-bearing planets which develop intelligent life; L is the 'life span' of the communicating technological culture.

Astronomy is crucial in source selection for targeted surveys. We assume that ETCs need billions of years to develop, so only stars with such lifetimes would be suitable candidates for SETI. These stars must be similar to our Sun. There are approximately 1,000 such stars within 100 light years from the Sun¹⁸. More than 50 extrasolar planetary systems are currently known, but most of these are unlike our Solar System. At present we can detect planets that have Saturn-like masses with orbits close to the star, but because of limitations to measurements, we still have not found an Earth-like planet orbiting a Sun-like star. Finding these will require hundreds of times more accuracy than is available now; such searches will be conducted from satellites in the next decades. The value of N remains highly uncertain. Even if we had a perfect knowledge of the first two terms in the equation, there are still five remaining terms, each of which could be uncertain by factors of 1,000.

interrupted by the interstellar medium between an ETC and the Earth²³, so an otherwise constant signal might be detectable only occasionally. It is simple to imagine ways round this problem, such as redundancy, repetition or the transmission of a second signal at twice or one-half the chosen wavelength²⁴. If the ETCs do not want to transmit signals deliberately, we may have to eavesdrop. One example of an unintended signal is the broadcast from a powerful television transmitter. But the reliable detection of unintended messages would require much larger telescopes, and in the case of television, the ETCs might use transmission via satellites or cable, in which the chances of finding broadcasts would decline still further.

Beyond radio

There is no reason why SETI should be restricted to radio wavelengths — it is possible that ETCs might transmit in other parts of the electromagnetic spectrum, such as the infrared or optical ranges^{25,26}. The advantage of the latter is simply one of energy density: with optical systems, very high ERPs can be obtained with modestly sized optical telescopes. Indeed, optical SETI (OSETI) is just now beginning with 1-m class optical telescopes (ref. 27 and Box 4). The OSETI searches for type II and III civilizations should be finished quickly, but for type I civilizations, OSETI will require much more time. If type I ETCs are common, one can expect success from a complete survey of stars to a distance of 100 light years. A targeted survey would require a few years. If, however, there is an ETC for every million Sun-like stars, the time will be much longer. Using the ATA in the centimetre wavelength range or the modern OSETI detector systems in the optical or infrared ranges, it would be possible to examine a number of stars simultaneously. Such techniques should shorten the time needed for such searches to a few decades.

ET intentions

If signals are transmitted with the intention that they are detected, we would expect that ETCs would want us to recognize their signals as

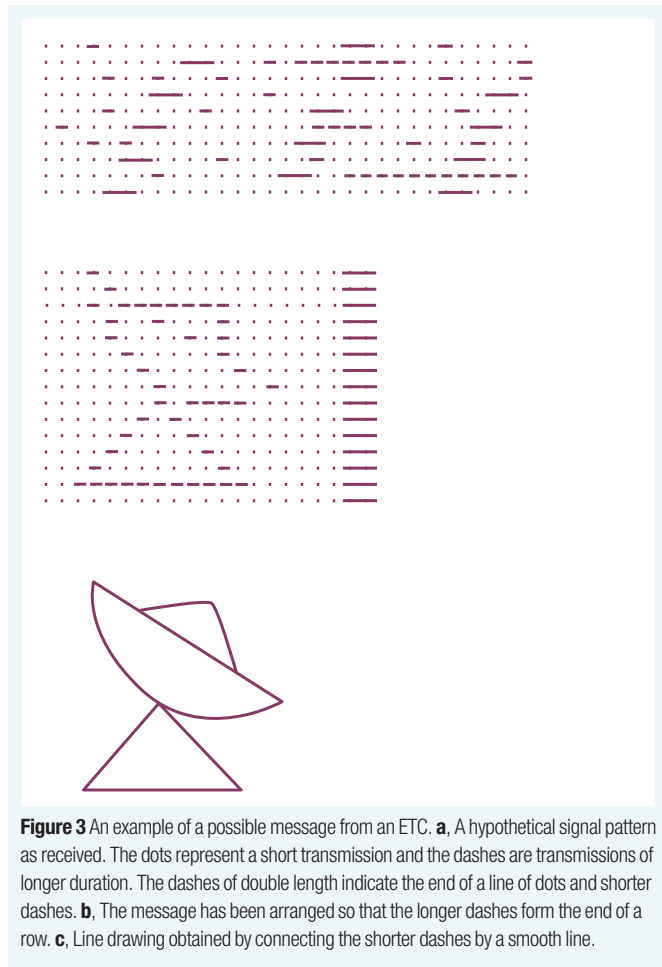


Figure 3 An example of a possible message from an ETC. **a**, A hypothetical signal pattern as received. The dots represent a short transmission and the dashes are transmissions of longer duration. The dashes of double length indicate the end of a line of dots and shorter dashes. **b**, The message has been arranged so that the longer dashes form the end of a row. **c**, Line drawing obtained by connecting the shorter dashes by a smooth line.

artificial. Broadcasts should have characteristics incommensurate with natural signals. The problem is that the characteristics of natural radio emissions vary over an extraordinarily large range. The narrowest spectral line has a width of 600 Hz, which is also a transition of the hydroxyl radical at 18 cm (ref. 14) in the 'water hole'. This line has a rather high intensity, but we are certain that it is natural as the noise statistics are typical of random processes and the time variability is very slow. At the other extreme are signals from pulsars, which show time variations on microsecond scales. At first, there was some speculation that pulsar signals were broadcasts from ETCs, but this was dismissed when measurements showed that the pulses were present over a wide frequency range. No ETC would be so wasteful, so the emission was taken to be natural. This emission has also been identified with rotating neutron stars¹⁴. In the near infrared, several stars were found to have large amounts of excess emission. Because such phenomena had already been discussed in terms of ETCs, this was taken as a signature of type II civilizations (ref. 28 and Box 2) in the course of shielding their parent star within a Dyson sphere. However, the radiation was found to be broadband, with only slow time variations, and these objects turned out to be dust-enshrouded stars¹⁴.

In summary, we assume that any ETCs would be aware of astronomical phenomena and would construct signals that will appear to be anything but normal. If ETCs send messages, these would be sent at specific wavelengths. They would have widths of a few hertz or pulses of nanosecond duration, but not both. How can we differentiate between terrestrial interference and ETC messages? If the signals from ETCs cover a large bandwidth, these would be subject to greater time delays at longer wavelengths in travelling through the slightly ionized medium of interstellar space¹⁴. The actual ETC messages at shorter wavelengths would arrive earlier. Terrestrial interference would show no such delay, so we would be able to differentiate between broadband

Box 4

Optical SETI

There is an advantage in transmitting signals at short wavelengths. This explains the interest in optical SETI (OSETI) in which searches are done at optical wavelengths, which are very much shorter than radio wavelengths. In OSETI, receiver noise floor is very much lower, increasing receiver sensitivity. For transmission, it is simple to show that a 1-m telescope, operating in the optical range, can produce one hundred million times the ERP as a 100-m radio telescope operating at 50 cm. In addition, systematic effects such as interference should be less in the optical or infrared than the radio wavelength region and the rate of information transfer is faster because of the larger bandwidths. Filters in the optical range are less selective than in the radio range, but an advantage is that ETCs could send messages using nanosecond pulses designed to look artificial and thus distinguishable from natural sources of electromagnetic radiation.

OSETI proponents also make an argument based on properties of the interstellar medium. Ionized clouds in the interstellar medium²³ scatter and absorb light much less than radio signals. However, optical or infrared signals are absorbed by dust in interstellar clouds. The targets for OSETI are selected on the basis of visible light or near-infrared measurements, and so would not be affected by intervening material. Thus the overall effect of interstellar clouds on optical or near-infrared communications is smaller.

The following example illustrates the advantages of OSETI in regard to effective radiated power²⁷. An ETC orbiting a Sun-like star could use a laser to illuminate a 1-m optical telescope through narrowband optical filters. The ETC could then produce a short pulse lasting a microsecond or less. This would produce a flash 300,000 times as bright as their Sun. Even without optical filtering, the flash would still be 30 times as bright as their Sun, and this factor would rise to 3,000 if the diameter of the telescope were increased to 10 m, as with the Keck telescope. Because of the short pulse length, such OSETI signals would not be found in conventional optical surveys.

ETC signals and local interference. For steady, narrowband signals, a fixed direction on the sky, modulation to transmit a message, and periodic Doppler shifts in wavelength caused by orbiting a star would be signs of extraterrestrial origin. However, short-lived, narrowband signals could be either interference from Earth or ETC signals.

What would be the content of an ETC signal? There have been many studies of the coding and content of hypothetical messages. The central problem of interpreting a message from an ETC is that we would have no idea about its language or syntax. The most reasonable communication mode would be mathematical and/or pictorial. Figure 3 shows a simple example of a possible ETC transmission, which consists of a picture from a rectangular array of '1's and 0's²². Because the time delay would be decades or even centuries, there could be no 'conversations', but a message announcing 'we are here' and presumably 'this is what we look like'.

Where are they?

If ETCs exist, they are not making their presence obvious. This in itself suggests that type III and perhaps type II civilizations are at best extremely rare. There are, however, many possible reasons why we have not made contact with ETCs. First, there may simply be very few⁵. Second, there may be a number of ETCs, but these may be sending messages in optical or near-infrared ranges that we have yet to explore comprehensively²⁵. Third, there may be ETCs, but these may not be interested in communicating²⁹ and choose to keep themselves hidden. This is more speculative since it depends on the cultural aspects of ETCs³⁰. From searches so far, the lack of contact shows us that transmissions, if any, are weak or intermittent signals (or both). The detection of ETC messages will take time and effort.

As the Drake equation (Box 3) shows, SETI is a business full of imponderables, of unknowable answers to impossible questions. We can, however, take comfort in some of the near-certainties of physics. For example, there seems to be no hope for faster-than-light travel, so actual visits from ETCs are unlikely. Even with the most efficient propulsion systems, the energy needed to reach stars at 10 light years in 20 years would be the equivalent of the present world consumption for 1,000 years (ref. 31). Such expenditure of energy would hardly deter a type III ETC, but even then, broadcasts make more energetic sense than personal appearances. There have been suggestions that ETCs might populate space with self-replicating machines in space probes³². This would allow colonization of large regions of space in relatively short intervals of time, but it seems vastly more complex than communicating by means of electromagnetic radiation⁶.

And what of ourselves? One can only speculate on the effect that contact with an ETC would have on humankind. Such a discussion exceeds the scope of this essay, which has been confined (deliberately) to technical matters. But the simple knowledge of the existence of ETCs would doubtless have far-reaching effects on our relationship with the Universe — as, no doubt, would be a persistent failure to detect ETCs, as this would confirm our uniqueness. □

- Hart, M. H. An explanation for the absence of extraterrestrials on Earth. *Q. J. R. Astron. Soc.* **16**, 128–135 (1975).
- Sagan, C. & Drake, F. D. The search for extraterrestrial intelligence. *Sci. Am.* **232**, 80–89 (May 1975).
- Vogt, S. S., Marcy, G. W., Butler, P. R. & Apps, K. Six new planets from the Keck Precision Velocity Survey. *Astrophys. J.* **536**, 902–914 (2000).
- McKay, D. S. et al. Search for past life on Mars: possible relic biogenic activity in martian meteorite ALH84001. *Science* **273**, 924–930 (1996).
- Ward, P. D. & Brownlee, D. *Rare Earth: Why Complex Life is Uncommon in the Universe* (Copernicus Books, New York, 2000).
- Cocconi, G. & Morrison, P. Searching for interstellar communications. *Nature* **184**, 844–846 (1959).
- Drake, F. D. in *Current Aspects of Exobiology* (eds Mamikunian, G. & Briggs, M. H.) 323–345 (Pergamon, New York, 1965).
- Baum, L. F. *The Wizard of Oz* (Bobbs-Merrill, New York, 1900).
- Tarter, J. in *The Search for Extraterrestrial Intelligence* (eds Kellermann, K. I. & Seilstad, G. A.) 79–98 (National Radio Astronomy Observatory, Greenbank, 1986).
- Tarter, J. & Klein, M. in *Progress in the Search for Extraterrestrial Life* (ed. Shostak, G. S.) *ASP Conf. Ser.* **74**, 457–469 (The Astronomical Society of the Pacific, 1995).
- McDonough, T. in *Progress in the Search for Extraterrestrial Life* (ed. Shostak, G. S.) *ASP Conf. Ser.* **74**, 419–422 (The Astronomical Society of the Pacific, 1995).
- Pierson, T. in *Progress in the Search for Extraterrestrial Life* (ed. Shostak, G. S.) *ASP Conf. Ser.* **74**, 443–444 (The Astronomical Society of the Pacific, 1995).
- Werthimer, D., Ng, D., Bowyer, S. & Donnelly, C. in *Progress in the Search for Extraterrestrial Life* (ed. Shostak, G. S.) *ASP Conf. Ser.* **74**, 293–301 (The Astronomical Society of the Pacific, 1995).
- Rohlf, K. & Wilson, T. L. *Tools of Radio Astronomy* 3rd edn (Springer, Heidelberg, 1999).
- Leigh, D. & Horowitz, P. in *Bioastronomy '99—A New Era in Bioastronomy* (eds Lemarchand, G. & Meech, K.) *ASP Conf. Ser.* **213**, 459–465 (The Astronomical Society of the Pacific, 2000).
- Welch, W. J. & Dreher, J. in *Radio Telescopes* (ed. Butcher, H. R.) *Proc. SPIE* **4015**, 8–18 (2000).
- Oliver, B. M. & Billingham, J. (eds) *Project Cyclops*. NASA CR-114445 (1973).
- Kuiper, T. B. H. & Morris, M. Searching for extraterrestrial civilizations. *Science* **196**, 616–621 (1977).
- Mauersberger, R. et al. SETI at the spin-flip line frequency of positronium. *Astron. Astrophys.* **306**, 141–144 (1996).
- Kardashev, N. S. Optimal wavelength region for communication with extraterrestrial intelligence— $\lambda = 1.5$ mm. *Nature* **278**, 28–30 (1979).
- Kardashev, N. S. Transmission of information by extraterrestrial civilizations. *Sov. Astron.* **8**, 217–221 (1964).
- LePage, A. J. Where they could hide. *Sci. Am.* **283**, 40 (2000).
- Cordes, J. M. & Sullivan, W. T. III in *Progress in the Search for Extraterrestrial Life* (ed. Shostak, G. S.) *ASP Conf. Ser.* **74**, 325–334 (The Astronomical Society of the Pacific, 1995).
- Cohen, C. & Charlton, D. in *Progress in the Search for Extraterrestrial Life* (ed. Shostak, G. S.) *ASP Conf. Ser.* **74**, 313–322 (The Astronomical Society of the Pacific, 1995).
- Townes, C. H. At what wavelength should we search for signals from extraterrestrial intelligence? *Proc. Natl Acad. Sci. USA* **80**, 1147–1151 (1983).
- Kingsley, S. A. in *The Search for Extraterrestrial Intelligence in the Optical Spectrum* Vol. II (ed. Kingsley, S. A. & Lemarchand, G. A.) *Proc. SPIE* **2704**, 102–116 (1996).
- Howard, A. et al. in *Bioastronomy '99—A New Era in Bioastronomy* (eds Lemarchand, G. & Meech, K.) *ASP Conf. Ser.* **213**, 545–552 (The Astronomical Society of the Pacific, 2000).
- Dyson, F. J. in *Perspectives in Modern Physics* (ed. Marshak, R. E.) 641 (Wiley, New York, 1966).
- Brin, G. D. The 'great silence': the controversy concerning extraterrestrial intelligent life. *Q. J. R. Astron. Soc.* **24**, 283–309 (1983).
- Morrison, P. M. in *Interstellar Communication: Scientific Perspectives* 168–186 (ed. Ponnamperna, C. & Cameron, A. G. W. (Houghton-Mifflin, Boston, 1974).
- Oliver, B. M. Efficient interstellar rocketry. *International Astronautical Federation IAA-87-606*, 1–6 (1987).
- Bracewell, R. N. *The Galactic Club* (Freeman, San Francisco, 1974).
- Sullivan, W. T. III, Brown, S. & Wetherill, C. Eavesdropping: the radio signature of the Earth. *Science* **199**, 377–388 (1978).
- Glaser, P. E. Power from the Sun: its future *Science* **162**, 857–861 (1968).
- Wilson, T. L. & Huettnerich, S. *Tools of Radio Astronomy: Problems and Solutions* p. 15 (Springer, Heidelberg, 2000).

Humans in space

Ronald J. White* & Maurice Avernert†

*National Space Biomedical Research Institute and Baylor College of Medicine, One Baylor Plaza, NA-425, Houston, Texas 77030, USA

†NASA Ames Research Center, Moffett Field, California 94035-1000, USA

Many successful space missions over the past 40 years have highlighted the advantages and necessity of humans in the exploration of space. But as space travel becomes ever more feasible in the twenty-first century, the health and safety of future space explorers will be paramount. In particular, understanding the risks posed by exposure to radiation and extended weightlessness will be crucial if humans are to travel far from Earth.

Accomplishments in engineering over the past 100 years have provided unprecedented opportunities for people to become mobile and travel rapidly on or near the surface of the Earth. Taking advantage of these opportunities, we have become citizens of the world, taking to the skies so often that the twentieth century will surely be known as the Century of Air Travel. However, even now the tools are in our hands to enable us to travel away from our home planet and become citizens of the Solar System. There are many reasons why human voyages of space exploration are worthy of serious discussion at the dawn of the new millennium. In fact, humans are beginning to develop the robust infrastructure that will make the twenty-first century the Century of Space Travel. But this bold step must be taken with due concern for the health and safety of future space explorers. This, in turn, means that we must develop both a new understanding of the risks posed by the potentially dangerous levels of radiation and extended weightlessness associated with future missions of exploration and a more effective means of coping with these risks. This article describes some of the benefits and risks of human missions of space exploration and summarizes the critical questions and issues that must be dealt with now, before fundamental decisions are reached concerning the appropriate time for humans to move away from Earth on voyages of exploration.

Benefits of human space exploration

Human explorers sent to live in space and to travel to other planetary bodies projects a captivating and alluring image. The glamour and excitement of human voyages, as opposed to automated, robotic missions, does much for public morale and for our need as humans to go where we have not yet ventured. But the reason for human exploration goes far beyond these emotional aspects and is in fact a necessary component of exploring the Universe¹⁻³.

The experiences of astronauts and cosmonauts over the past four decades have proven the merits and necessity of humans as space explorers. Complex tasks, scientific experimentation, and repair and troubleshooting of equipment and hardware, for example, all require human capabilities and judgement. There are myriad examples of humans being required for the success of mission discoveries. The initially flawed Hubble Space Telescope is a case in point. Astronauts repaired the faulty scope and have been required for continued servicing of this multimillion-dollar project. Without human intervention the project would have a fraction of the value NASA has been able to glean. The

Apollo lunar missions provide another example. Astronauts were imperative on the lunar surface for remedying unforeseen problems, such as repairing the rover vehicle and using their training and communication with ground-based scientists to select representative samples from a given lunar location and recognize and evaluate interesting findings. Humans will be imperative for similar goals on the surface of Mars.

A principal goal of space travel is planetary exploration and the search for life³. In-depth understanding of planetary histories and processes will require field investigation. Although automated robotic missions are often touted as the most efficient and cost-effective way to perform such investigation, current robots are inadequate for accomplishing the complex iterative processes required for successful scientific field studies. As outlined in a recent report², a human mission has much greater promise for answering principal strategic questions than does a larger number of robotic missions; it also creates many more options for modes of exploration that cannot be achieved robotically. This in itself makes human missions more cost effective on scientific grounds.

Human presence will be required for the reasoning and responses necessary to accommodate unexpected discoveries and to perform real-time testing of hypotheses. Evidence of life, for example, is likely to be hidden and microscopic, requiring long-distance travel over rugged terrain, digging to great depths, surveying numerous sites, and finely dissecting rock and soil layers. These are all intricate tasks that far exceed the current capabilities of robots and are likely to for a long time. The recent Mars Pathfinder mission highlights this point. Although this was a successful mission for achieving its goals, the limitations of the Pathfinder's mobile Sojourner rover for scientific investigation were vast. The rover was able to travel just over 100 metres around the landing site before communication was terminated, which limited scientific return severely. Chemical analysis of some of the rock samples by the onboard spectrometer revealed an unexpected composition. However, without more sophisticated intervention it was impossible to perform further evaluation, and without such information the scientific value was extremely limited. An astronaut scientist would have been able to make a field identification of this rock, collect samples and perform field tests, which would have untold scientific value. There is great value in *in situ* analysis of samples on planetary surfaces. Such sophisticated analysis would be very difficult if not impossible to automate in the foreseeable future and furthermore would not lend to iterative experimentation arising from the results found.

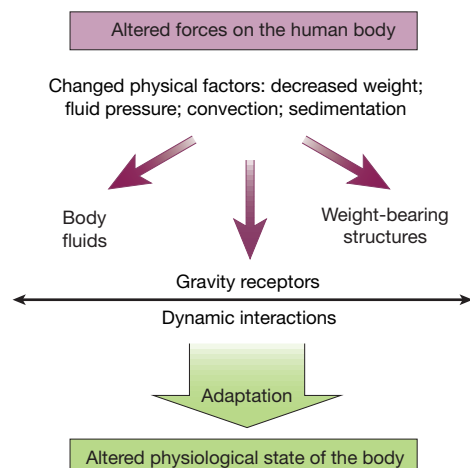


Figure 1 The general effects of space flight or reduced gravity (for example, on the surface of Mars) on the physical and biological elements of the human body.

Answering the principal scientific questions outlined for exploration of Mars will require investigation in geology, palaeontology, biology, chemistry, geophysics, atmospheric science and climatology³. Initial investigations and assessments in these areas can be achieved by the use of robots, but follow-up detailed investigation and discovery will have to be done on the surface of the planet by human crews. On balance, human abilities and capabilities are critical for deriving the maximum benefits of missions of exploration.

Risks and challenges of human space exploration

Voyages of exploration will subject space travellers to three serious and related challenges: (1) changes in the physical forces on and within the body brought about by a reduction in weight of the body's components; (2) psychosocial changes induced by the long-term confinement of such a voyage without the possibility of escape; and (3) changes in the levels and types of radiation in the environment. These changes, which act simultaneously, precipitate a cascade of time-related events in the human body about which we have been learning slowly for the past 40 years⁴. The integrated and unmitigated responses of the body to these challenges present real risks to the health of the humans undertaking such missions and to the satisfactory completion of the missions themselves. Some of the risks pose a greater threat than others do, and the level of understanding of the physiological responses to space flight varies depending on the body system in question. Fortunately, it seems that most of these risks may be reduced to an acceptable level through a vigorous research programme.

Changes in physical forces

Figure 1 illustrates the general biological consequences of the primary physical events occurring during space flight. Because weight is decreased to very nearly zero for much of a mission, the weight-bearing structures of the body are subject to a different set of stresses. Changed hydrostatic pressure gradients along the body axes cause a fluid shift within the body, and the input to the body's many gravity receptors is altered significantly. Almost all of the body's components and systems participate in the response to these events. Ultimately, many of these systems seem to adapt, but questions still remain concerning the extent of that adaptation and the stability of the resultant state of the body. Because of the brevity of this review, only one example of the effect of weightlessness will be given. For a more complete list of the risks and the physiological systems implicated by these risks, see <http://criticalpath.jsc.nasa.gov/>.

Prolonged exposure to weightlessness during extended space flight (for example, voyages lasting a year or more) will significantly

increase the risk of fracture⁵. Measurements of bone mineral density during the Mir space flights of 4.5–14.5 months indicate an average loss of $5.6 \pm 0.8\%$ from the lumbar spine, $11 \pm 1.4\%$ from the pelvis and $8.1 \pm 1.2\%$ from the proximal femur⁶. The extent of bone loss for individual astronauts or cosmonauts is considerable, varying from 0% to up to 20% (ref. 7). This stands in sharp contrast to a decrease in bone mass of 2–3% per decade in postmenopausal women. Although the mechanisms promoting this variation in individual rates of bone loss are undetermined, it is possible that once weightlessness occurs, the astronauts' genetic background and the actions of locally acting cytokines shift bone remodelling to favour an increase in bone resorption. Most observers agree that bone loss is likely to be progressive, at least to the point that fracture poses an immediate risk during an extended space flight, such as a proposed 3.5-year exploration mission to Mars.

In terms of countermeasures to bone loss, various exercise regimens have not proven effective in space nor has exercise significantly aided the rapid re-establishment of bone mass after return to Earth. Dietary calcium and vitamin D supplements have also not prevented bone loss⁸. However, we know that bisphosphonates limit bone loss effectively where bone resorption is increased, and bed-rest studies examining this approach are underway. Newer agents, perhaps derived from other current studies on regulatory factors, may prove more effective ultimately. One hypothesis is that both resistive exercise and a pharmacological agent will be required to prevent bone loss during extended space flight.

Some of the other effects of weightless space flight include: cardiovascular and fluid-related problems of orthostatic hypotension immediately following space flight^{9–11}, the possibility of altered cardiac susceptibility to ventricular arrhythmias¹², and reduced cardiac muscle mass and diminished cardiac function¹³; neurovestibular-related problems at the beginning of a flight involving space motion sickness¹⁴, and during and just after landing involving disorientation, gait changes, and impaired balance and neuromuscular coordination^{15–18}; muscle-related problems of atrophy involving loss of muscle mass, strength and endurance^{19–22}; circadian rhythm-related problems involving sleep and performance^{23,24}; and immune-related problems involving infections and immunodeficiency^{25,26}.

Psychosocial and neurobehavioural changes

Based on documented evidence from both US and Russian space missions in which astronauts and cosmonauts experienced personal and interpersonal problems^{27,28}, adverse psychosocial reactions

Table 1 Psychosocial and neurobehavioural issues under investigation for improving crew health and safety during prolonged space flight

Biological mechanisms of dysfunction	What cellular, molecular and organismic changes in the function of the nervous system occur during conditions present in prolonged space flight?
Motivation, cognition and performance	How can neurobehavioural and cognitive functions, physical and emotional distress, motivation and operational performance be assessed and enhanced during a mission?
Pharmacology in space	What does microgravity do to the pharmacokinetics, bioavailability, side effects and efficacy of medications for stress reactions?
Individual factors	What characteristics of astronauts are predictive of neurobehavioural resiliency and stable interpersonal performance during a mission?
Team and interpersonal optimization	What aspects of leadership, crew selection and composition, training, communication, social interaction, decision-making and error management facilitate functioning within a multicultural crew and between the crew and ground personnel?
Organizational, cultural and management factors	What organizational, professional and management goals, policies and priorities most affect crew communication, performance, problem solving and, ultimately, health and safety?

The six interrelated themes and example questions define the range of factors that are critical for improving crew health and safety during prolonged space flight and for optimizing individual and crew performance.

Table 2 A comparison of some space-related health concerns with medical issues on Earth

Research area	Space	Earth
Bone	Bone loss and increased fracture risk Increased kidney-stone formation Injury to soft connective tissue	Osteoporosis and other bone disorders
Cardiovascular	Postflight orthostatic intolerance Cardiac atrophy Heart rhythm disturbances	Orthostatic intolerance Heart disorders, such as sudden cardiac death due to heart rhythm disturbances
Performance/Sleep	Errors due to sleep loss and disruption of the biological clock	Sleep problems due to jet lag, shift work and extended work schedules Accidents due to sleepiness
Immunology/Infection	Activation of dormant viruses in the body Increased infection risk Space flight-related anaemia Interference with wound healing	Immune system disorders Viral outbreaks due to stress conditions (shingles, cold sores) Anaemia and other blood disorders
Muscle	Muscle loss and atrophy	Muscle wasting diseases Muscle weakening due to prolonged bed rest, immobilization, nerve crush injury and ageing
Neurovestibular	Space motion sickness and body orientation problems Re-entry vertigo Postflight dizziness, balance, posture and gaze stability	Vertigo and other balance disorders
Radiation effects	Cancer Damage to central nervous system Cataracts and other diseases	Risks from exposure to naturally occurring and work-related radiation

among astronauts during prolonged flights are now recognized as a serious risk to mission success. Astronauts aboard exploration-class space missions will endure behavioural challenges for a much longer period of time and in different circumstances. Exploration mission stressors include confinement for up to three years with the same small group of people; isolation from family and friends; limited communication with Earth, including a delay of up to 24 minutes in bi-directional communications; and loss of privacy due to habitability constraints. Additional neurobehavioural risks are posed by prolonged exposure to microgravity, radiation and equipment failure in space. Judging from current evidence, language, culture, gender and differences in work role will also pose challenges to crew communication and effectiveness.

Without mitigation, these stressors can impose a burden on astronaut behavioural capability and health, both individually and collectively. They have the potential to erode cognitive performance; change neuroendocrine, cardiovascular and immune responses; disrupt appetite, sleep and other basic regulatory physiology; lead to neuropsychiatric impairment through anxiety and depression; and potentiate serious interpersonal problems among crewmembers. Thus, research is focusing on finding ways to ensure that astronaut neurobehavioural health is maintained, that performance capability is facilitated by appropriate habitat and human-systems interfaces, and that crew psychosocial functioning is optimized effectively.

Today, nearly all of this research is done on Earth in laboratory or special environments (such as Antarctica) and focuses on psychological (for example, personality) and behavioural (for example, leadership) characteristics of individuals and groups in relation to performance and stress reactions^{29,30}. Objective measures of neurobehavioural performance are being investigated along with a number of new unobtrusive technologies for computer recognition of emotional distress. The neurobiological processes underlying stress and arousal responses are also being studied in animals and human models to identify the most appropriate behavioural and pharmacological countermeasures. The ultimate goal is to reduce risk by appropriate monitoring of physiological function and behaviour, and by having an appropriate arsenal of countermeasures available to enhance performance, motivation and the quality of life during an exploration voyage. Table 1 provides examples of the scope of issues under scientific investigation.

Changes in the radiation environment

Human missions of exploration will expose crewmembers to transient radiation from solar particle events and to continuous radiation from high-energy galactic cosmic rays^{31,32}. The protons and high-atomic-number energetic particles (HZE) involved may exert

sizeable biological effects even at low fluence, and there are considerable uncertainties associated with secondary particle effects (for example, HZE fragments or neutrons). Although the health risks from exposure to radiation (X-rays, gamma rays or electrons) encountered on Earth are comparatively well known, the health risks from space radiation remain far from understood. Several independent factors contribute to the overall risk to astronauts exposed to the complex radiation environment of exploration missions. Of primary concern is the induction of late-occurring cancers³³. But damage to the central nervous system is also potentially a mission-compromising event because of the possibility of cell loss from radiation damage affecting the functional integrity of the central nervous system³⁴. Recent studies^{35,36} also point to previously unknown mechanisms of radiation-induced cellular pathologies based on the communication between damaged and undamaged cells and the induction of unstable states that lead to late expression of genetic damage. Space radiation seems to be uniquely effective in causing such cellular changes. Current research strategies focus on shielding, risk estimation and risk mitigation and use both cellular systems and animal models, together with proton and HZE particle accelerators.

Future directions — the challenge of integrative physiology

The experiences of the past 40 years of space-related research and health care have shown that, to manage effectively the health and related mission risks of future space explorers, it is not enough to separately address the loss of bone that occurs during a nearly weightless state, with its attendant increased risk of fracture, or the problems of increased cancer risk caused by the natural radiation that accompanies space flights away from low-Earth orbit. Meeting the health-related challenges of human space exploration requires that one abandon any model of the human body that has the muscles, bones, heart and brain acting independently. Body parts will not travel on exploration missions. Instead, the individual space traveller's body must be viewed realistically, with all parts connected and fully interacting. Development and use of such an integrative approach must capitalize on the investments that have been and continue to be made in molecular biology and on the new and emerging capabilities in computing, information storage, modelling, and fast, parallel processing that characterize today's technology. This will not be easy; the problems and challenges that must be faced are many and great and have been discussed for several years³⁷.

In the United States, the National Space Biomedical Research Institute has already embarked on just such an integrative physiology programme with NASA support (see www.nsbri.org). The ultimate result of this long-term programme will be the development of a

quantitative description of a healthy human being that contains state-of-the-art information on each component of the body and on how these components relate to each other. This description, termed a 'digital human', will contain virtually everything known about human physiology, from biochemical to cellular to organ to system information and then to interactions among the systems. It will serve as an integrated repository of knowledge on mechanisms and their coordinated operation in the intact human.

Realizing this goal involves the demonstration of a deep understanding of functional elements in human performance, from systems physiology to the individual genotype responsible ultimately for performance characteristics. It involves the development of models of component systems assembled in a hierarchical or relational way, and the development of an understanding of how environmental stresses impact on the function and adaptation of the resultant phenotype. In practical terms, realizing this goal also requires a strategy for integrating components and results from a number of investigators (and laboratories) into a coherent synthesis of human function.

For the space programme, the goal of such work is to provide a personalized human model for each member of the crew of an exploration voyage. Comprehensive individual models of the anatomy, physiology, functional status, and medical and environmental history of each astronaut will then contribute to monitoring, diagnosis, treatment and outcome prediction, as well as assisting mission planners and the crew themselves in reducing health and mission risks. But the payoff from this integrative approach, if successful, will extend far beyond the world of space exploration. An understanding of the functioning of the healthy body will enable researchers to probe not only the mechanisms responsible for the many changes that occur during space flight, but also those factors responsible when some pathway or component within the body becomes dysfunctional, as it does during injury or disease. In fact, as Table 2 shows, many of the medical concerns associated with space flight are related strongly to familiar medical issues on Earth. The problem of providing appropriate health care for future space explorers acts only as a focusing lens to enable us to see more clearly the problems we face in providing quality health care in general. The great challenge of biology is to develop the tools to implode the explosive accumulation of knowledge and information about our biological selves. Perhaps, then, the ultimate reason for human space exploration is to enable us to discover ourselves. □

1. Hoffman, S. J. & Kaplan, D. L. (eds) *Human Exploration of Mars: The Reference Mission of the NASA Mars Exploration Study Team* <<http://www-sn.jsc.nasa.gov/marsref/contents.html>> (Lyndon B. Johnson Space Center, Houston, TX, 1997).
2. Space Studies Board, National Research Council. *Scientific Opportunities in the Human Exploration of Space* (National Academy Press, Washington DC, 1994).
3. Special Report: Sending Astronauts to Mars <<http://www.sciam.com/2000/0300issue/0300currentissue.html>> *Sci. Am.* **97** (March 2000).
4. Nicogossian, A., Huntoon, C. & Pool, S. (eds) *Space Physiology and Medicine* 3rd edn (Lea and Febiger, Philadelphia, 1994).
5. Morey-Holton, E. R., Whalen, R. T., Arnaud, S. B. & Van Der Meulen, M. C. in *Handbook of Physiology: Section 4: Environmental Physiology* (eds Fregly, M. J. & Blatteis, C. M.) 691–719 (American Physiological Society, New York, 1996).
6. Grigoriev, A. I. *et al.* Clinical and physiological evaluation of bone changes among astronauts after long-term space flights. *Aviakosm Ekolog Med.* **32**, 21–25 (1998). [In Russian.]
7. Vico, L. *et al.* Effects of long-term microgravity exposure on cancellous and cortical weight-bearing bones of cosmonauts. *Lancet* **355**, 1607–1611 (2000).
8. Smith, S. M. *et al.* Calcium metabolism before, during and after a 3-month spaceflight: kinetic and biochemical changes. *Am. J. Physiol.* **277**, R1–R10 (1999).
9. Buckley, J. C. *et al.* Orthostatic intolerance following spaceflight. *J. Appl. Physiol.* **81**, 7–18 (1996).
10. Wautenpaugh, D. E. & Hargens, A. R. in *Handbook of Physiology: Section 4: Environmental Physiology* (eds Fregly, M. J. & Blatteis, C. M.) 631–674 (American Physiological Society, New York, 1996).
11. Reyes, C., Freeman-Perez, S. & Fritsch-Yelle, J. Orthostatic intolerance following short and long duration spaceflight. *FASEB J.* **13**, A1048 (1999).
12. Fritsch-Yelle, J. M. *et al.* An episode of ventricular tachycardia during long-duration spaceflight. *Am. J. Cardiol.* **81**, 1391–1392 (1998).
13. Levine, B. D., Zuckerman, J. H. & Pawelczyk, J. A. Cardiac atrophy after bed-rest deconditioning: a non-neural mechanism for orthostatic intolerance. *Circulation* **96**, 517–525 (1997).
14. Oman, C. M., Lichtenberg, B. K., Money, K. E. & McCoy, R. K. MIT/Canadian vestibular experiments on the Spacelab-1 mission: space motion sickness: symptoms, stimuli, and predictability. *Exp. Brain Res.* **64**, 316–334 (1986).
15. Lackner, J. R. & DiZio, P. Human orientation and movement control in weightless and artificial gravity environments. *Exp. Brain Res.* **130**, 2–26 (2000).
16. Reschke, M. F., Bloomberg, J. J., Harm, D. K. & Parker, D. E. in *Space Physiology and Medicine* 3rd edn (eds Nicogossian, A. E., Huntoon, C. L. & Pool, S. L.) 261–285 (Lea and Febiger, Philadelphia, 1994).
17. Young, L. R., Oman, C. M., Watt, D. G. D., Money, K. E. & Lichtenberg, B. K. Spatial orientation in weightlessness and readaptation to Earth's gravity. *Science* **225**, 205–208 (1984).
18. Dauntun, N. G. in *Handbook of Physiology: Section 4: Environmental Physiology* (eds Fregly, M. J. & Blatteis, C. M.) 765–783 (American Physiological Society, New York, 1996).
19. Baldwin, K. M. Effect of spaceflight on the functional, biochemical, and metabolic properties of skeletal muscle. *Med. Sci. Sports Exerc.* **28**, 983–987 (1996).
20. Edgerton, V. R. *et al.* Human fiber size and enzymatic properties after 5 and 11 days of spaceflight. *J. Appl. Physiol.* **78**, 1733–1739 (1995).
21. Fitts, R. H., Riley, D. R. & Widrick, J. J. Physiology of a microgravity environment. Invited review: Microgravity and skeletal muscle. *J. Appl. Physiol.* **89**, 823–839 (2000).
22. Edgerton, V. R. & Roy, R. R. in *Handbook of Physiology: Section 4: Environmental Physiology* (eds Fregly, M. J. & Blatteis, C. M.) 721–763 (American Physiological Society, New York, 1996).
23. Czeisler, C. A. & Khalsa, S. B. in *Principles and Practice of Sleep Medicine* 3rd edn (eds Kryger, M. H., Roth, T. & Dement, W. C.) 353–375 (Saunders, Philadelphia, 2000).
24. Gündel, A., Polyakov, V. V. & Zulley, J. The alteration of human sleep and circadian rhythms during spaceflight. *J. Sleep Res.* **6**, 1–8 (1997).
25. Taylor, G. R. Overview of spaceflight immunology studies. *J. Leukoc. Biol.* **54**, 179–188 (1993).
26. Gmünder, F. K. & Cogoli, A. in *Handbook of Physiology: Section 4: Environmental Physiology* (eds Fregly, M. J. & Blatteis, C. M.) 799–814 (American Physiological Society, New York, 1996).
27. Kanas, N. Psychiatric issues affecting long duration space missions. *Aviat. Space Environ. Med.* **69**, 1211–1216 (1998).
28. Ellis, S. R. Collision in space. *Ergon. Design* **8**, 4–9 (2000).
29. Connors, M. M., Harrison, A. A. & Akins, F. R. *Living Aloft* (NASA, Washington, 1985).
30. Palinkas, L., Gunderson, E. K. E., Holland, A. W., Miller, C. & Johnson, J. C. Predictors of behavior and performance in extreme environments: the Antarctic space analogue program. *Aviat. Space Environ. Med.* **71**, 619–625 (2000).
31. Schimmerling, W. Radiobiological problems in space: an overview. *Radiat. Environ. Biophys.* **31**, 197–203 (1992).
32. Nelson, G. A. in *Handbook of Physiology: Section 4: Environmental Physiology* (eds Fregly, M. J. & Blatteis, C. M.) 785–798 (American Physiological Society, New York, 1996).
33. Space Studies Board, National Research Council. *Radiation Hazards to Crews of Interplanetary Missions* (National Academy Press, Washington DC, 1996).
34. Vazquez, M. E. Neurobiological problems in long-term deep space flights. *Adv. Space Res.* **22**, 171–183 (1998).
35. Azzam, E. I., de Toledo, S. M. & Little, J. B. Direct evidence for the participation of gap junction-mediated intercellular communication in the transmission of damage signals from α -particle irradiated to nonirradiated cells. *Proc. Natl Acad. Sci. USA* **98**, 473–478 (2001).
36. Kadhim, M. A. *et al.* Transmission of chromosomal instability after plutonium alpha-particle irradiation. *Nature* **355**, 738–740 (1992).
37. Boyd, C. A. R. & Noble, D. (eds) *The Logic of Life: The Challenge of Integrative Physiology* (Oxford Univ. Press, Oxford, 1993).

Acknowledgements

The authors acknowledge many helpful discussions with R. J. Cohen, J. F. Dicello, D. F. Dinges, C. Golden, A. R. Kennedy, J. I. Leonard, C. M. Oman, R. J. Schwartz and J. R. Shapiro. Support for the preparation of this paper was provided by NASA, both directly and through Cooperative Agreement NCC 9-58 with the National Space Biomedical Research Institute.

Where are the dolphins?

Jack Cohen & Ian Stewart

Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK

Interest in extraterrestrial life has tended to focus on a search for extrasolar planets similar to the Earth. But what of forms of intelligent life that are very different from those found on Earth? Some features of life will not be peculiar to our planet, and alien life will resemble ours in such universals. But if intelligent, non-humanoid aliens exist, where might they be? Would they wish to visit Earth and would we know if they did?

Science currently knows of only one life-bearing world, but our sample is biased, because it is the world we live on. As we learn more about other regions of the cosmos, the prospects for Earth-like aliens seem ever more encouraging: there should be many places in the Universe that are very similar to planet Earth. Current scientific interest in extraterrestrial life is mostly a search for extrasolar planets similar to our own¹. The main exception is the ocean now thought to exist beneath Europa's icy surface^{2–4}, but even there the interest lies in the resemblances between this ocean and its terrestrial equivalents.

A more interesting question, however, is the possibility of aliens, especially intelligent ones, that are not like us: which is, after all, what 'alien' means. It is possible to imagine the existence of forms of life very different from those found on Earth, occupying habitats that are unsuitable for our kind of life. Some of those aliens might be intelligent and technological, because technology is an autocatalytic process⁵. It follows that some aliens might possess technology well in advance of our own, including interstellar transportation. So much is clear, but this train of logic begs the obvious question of where these intelligent, non-humanoid aliens might be. Where, then, are the dolphins?

Part of the answer is that the question is too parochial in its outlook. Dolphins are the nearest thing to intelligent aliens on this planet, but they are our close evolutionary cousins, and they share many of our own accidental features. There might, perhaps, be dolphin-like aliens, but the dolphin habitat as found in Earth's oceans may not be sufficiently conducive to the development of technology. Nonetheless, we cannot escape the big question⁶, raised in 1950 by Enrico Fermi: if intelligent aliens exist, why aren't they here?

Canonical answers⁷ to Fermi's question (henceforth 'alien' will imply intelligence unless otherwise stated) include:

- There are no aliens, and there never have been. Humanity is unique in the Universe.
- There have been plenty of aliens, but civilizations only moderately more advanced than ours always blow themselves up in nuclear wars.
- The lifespan of an alien civilization is only a few million years. They visited us ten million years ago, and will turn up again in ten million years' time, but there is nobody around right at the moment.
- Aliens exist, but interstellar travel is impossible because of relativistic limits on the speed of light, or because living creatures cannot survive it.
- Aliens exist, but are not interested in interstellar travel.
- Aliens exist and have interstellar travel, but they are not interested in contacting us⁸.

- Aliens exist, but galactic law forbids any contact with us because we are too primitive⁹ or violent¹⁰.
- Some aliens see it as their duty to eliminate all other forms of life that come to their attention. Any technological civilization will develop radio and TV, attract their attention, and be eliminated¹¹. They are on their way now.
- They are here already (the preferred answer on the Internet's UFO pages).

The evidence for the last assertion, as for the others, is poor. Eyewitness accounts of alien abductions are unconvincing, even when offered in good faith. One of us (J.C.) was on a radio programme with a woman who maintained that aliens had abducted her and stolen her baby. J.C. asked a pertinent question that had eluded everyone else: "Were you pregnant?" Her reply: "no".

Even if we consider, for the sake of argument, that aliens walk among us, we can assume that they are highly intelligent creatures from a technologically advanced civilization and not likely to be swanning around in gigantic machines, kidnapping the natives, or doing weird things to the natives' reproductive organs.

Xeno's paradise

The subject area to which this discussion belongs is often called astrobiology, although in science-fiction circles (where the topic has arguably been thought through more carefully than it has been in academic ones) the term 'xenobiology' is favoured. The difference is significant. Astrobiology is a mixture of astronomy and biology, and the tendency is to assume that it must be assembled from contemporary astronomy and biology. In contrast, xenobiology is the biology of the strange, and the name inevitably involves the idea of extending contemporary biology into new, alien realms.

Upon what science should xenobiology be based? The history of science indicates that any discussion of alien life will be misleading if it is based on the presumption that contemporary science is the ultimate in human understanding. Consider the position of science a century ago. We believed then that we inhabited a newtonian clockwork Universe with absolute space and absolute time; that time was independent of space; that both were of infinite extent; and that the Universe had always existed, always would exist, and was essentially static. We knew about the cell, but we had a strong feeling that life possessed properties that could not be reduced to conventional physics; we had barely begun to appreciate the role of natural selection in evolution; and we had no idea about genetics beyond mendelian numerical patterns. Our technology was equally primitive: cars were inferior to the horse, and there was no radio, television, computers, biotechnology or mobile phones. Space travel was the

stuff of fantasy. If the past is any guide, then almost everything we now think we know will be substantially qualified or proven wrong within the next 25 years, let alone another century. Biology, in particular, will not persist in its current primitive form. Right now, it is at a stage that is roughly analogous to physics when Newton discovered his law of gravity. There is an awfully long way to go.

Xenobiology seems unusual, because it will require a science of what might happen in addition to the science of what we know. However, many scientific explanations involve contemplating possibilities that do not occur in addition to those that do, so the novelty is less than it seems. (The concept of stability, for example, involves answering a 'what if' question: 'what would the system do if it was perturbed?') The concept of phase space provides a useful framework for such deliberations¹². The phase space of a system is the set of all conceivable states of that system, often equipped with a topological structure, in which states that differ only slightly are considered to be neighbours. DNA-space, for example, comprises all conceivable DNA sequences, whereas phenotypic space comprises all conceivable designs for organisms. Xenobiology is an exploration of xenospace, the space of possible aliens, together with alien evolutions, alien cultures, and other associated influences from context or content.

Rockets and space elevators

It is important not to let the science of what we do not know be over-constrained by the science of what we do know, or think we know. In particular, life is an emergent phenomenon^{5,12} that the Universe 'invented' as it developed. How big is nature's palette? We suspect it is much larger than most people imagine.

Physics is a poor guide here. The spectra of distant stars tell us that physics and chemistry elsewhere in the Universe follow the same principles that they do here. This belief is probably fairly accurate, if only because physics and chemistry are partly invented (human beings choose what contexts to place them in, and those contexts tend to be simple laboratory-based ones, not the 'wild' physics of the real Universe). This leads us to expect biology to be the same everywhere, too. But, even within Earth-like biology, the combinatorial possibilities of carbon compounds compromises this line of argument. Chemists have believed the physicists' claim that chemistry is reducible to physics, but the chemistry in stellar interiors, for example, may not be so reducible in any meaningful way. (We do not dispute that the chemistry in stars is a consequence of physical laws, but it is an emergent consequence, so the laws provide few useful insights.)

Similarly, biology is an emergent consequence of physics and chemistry, making it



incomprehensible in terms of the 'tame' physics of the laboratory. This is an appropriate place to introduce two contrasting images: the rocket and the space elevator¹³. Physics places an apparently unbreakable limit on the amount of energy needed to place a human being in orbit: the difference in gravitational potential of an object in orbit compared with that at ground level. The law of conservation of energy implies that it will never be possible to put a human being into orbit cheaply. This argument may seem flawless, but it assumes implicitly a particular context: that the sole traffic is upwards. Instead, consider the space elevator, a cable suspended from a geosynchronous satellite^{14–16}. It will be expensive to build, but once it exists one could ride into space very cheaply, powered by minerals from the asteroid belt coming down the elevator for human consumption. The space elevator does not violate the law of conservation of energy, but it demonstrates that in this context that law is irrelevant to cost. Indeed, energy limitations will soon cease to constrain human activities, just as memory limitations constrain our computations less than they once did.

The kind of chemistry understood by contemporary molecular biology is analogous to the rocket; but cells have been using space-elevator chemistry for aeons, which is

why life is such an effective trick. Biology results from chemistry that has been corrupted by evolution, and evolution on Earth has been going for at least 3.8 billion years (see review in this issue by Nisbet and Sleep, pages 1083–1091). This is deep time — too deep for scenarios expressed in human terms to make much sense¹⁷. A hundred years is the blink of an eye compared with the time that humans have existed on Earth. The lifespan of the human race is similarly short when compared with the time that life has existed on Earth. It is ridiculous to imagine that somehow, in a single century of human development, we have suddenly worked out the truth about life. After all, we do not really understand how a light switch works at a fundamental level, let alone a mitochondrion.

For similar reasons, it is probably pointless to search the heavens for radio signals from other worlds, as the Search for Extraterrestrial Intelligence (SETI) project aims to do (see refs 18, 19 and the review in this issue by Wilson, pages 1110–1114). It would be equally sensible to look for smoke signals. Radio did not exist on this planet a hundred years ago, and might become obsolete. If aliens communicate at all, they might use media as yet undiscovered by human technology. Even if radio were their medium of

choice, they might have encoded their transmissions for optimal efficiency. Moore²⁰ has shown that an optimally efficient coded message will be indistinguishable from black-body radiation. Imagine a Second World War radio operator picking up one of today's encrypted satellite TV channels: it would sound like static. Is this the true meaning of the cosmic background radiation?

What is life?

An essential component of xenobiology will be a reassessment of the nature of life. The current belief that DNA holds the key to life as a general phenomenon might reflect an unnecessarily narrow perspective. For example, it has been suggested that the concept of the 'gene' might soon be redundant²¹. From a xenobiologist's viewpoint, the problem with life on Earth is that it is a very limited sample, even of DNA-based organisms. DNA space contains about $10^{1,000,000,000}$ different sequences of comparable length to the human genome. Most of those sequences cannot occur in viable organisms, but even if we eliminate an overwhelmingly large fraction we are still left with, say, $10^{1,000,000}$ viable sequences. There are, perhaps, 10^7 – 10^8 species on the planet today. Although these numbers are the roughest approximations, they are sufficient to make the point — that the phase space of the possible is far greater than is realized by the actual. From this it follows that the detailed genetic constitution of life on Earth is an accidental result of local history, and not the inevitable conclusion of fate.

However, despite their seemingly limited diversity, Earth's current life-forms may be more typical in other, more important, ways, such as their relationship with their context. 'Life' is a name we give to certain emergent processes of complex systems^{5,22}. Until quite recently we used the word as a catch-all to cover anything on this planet that seemed to have some kind of individual autonomy. It then became evident that everything of that kind was using the same trick — DNA (or RNA) and associated biochemistry. We have therefore assumed that DNA is the sole route to autonomy and self-complication. However, the prevalence of the DNA mechanism on this planet may be just a historical accident. When any one such trick evolves, it quickly dominates — the trick, by its nature, is self-copying, and tends to swamp the competition.

None of this implies that alternatives, especially radical ones, cannot exist. For xenobiological purposes the answer to 'what is life?' cannot be a catalogue of DNA bases. It must involve the recognition that the abstract processes of life possess certain universal features, and that those features might have a large number of possible different physicochemical realizations.

Parochials and universals

Even on Earth, our view of what life is and where it can survive has changed considerably in recent years. Extremophiles survive in environments that would be lethal to humans (refs 23–29, and see the review in this issue by Rothschild and Mancinelli, pages 1092–1101). This suggests that we should not place too much reliance on alleged limitations of living organisms. But our evolution story, even ignoring extremophiles, hints at principles that might also apply to life more generally (see the review in this issue by Carroll, pages 1102–1109). And evolution itself is one such principle: it will apply to aliens as much as to us. Therefore some features of life on Earth will not be peculiar to our planet.

The key distinction lies between features that are 'universal' and those that are merely 'parochial'³⁰. The best current test for universality is to ask whether a feature of interest arose more than once, independently, in evolution on Earth. If the answer is yes, as it is for flight, photosynthesis, locomotion, limbs and predation, then the feature is a universal. If not, as for pentadactyl limbs in tetrapods, the feature is a parochial. Alien evolution will resemble ours in universals, but not in parochials. Many disputes about alien life stem from disagreements about which features are universal and which are parochial. Because it is all we know, it is easy to assume that carbon-based molecular structure, genetics based on DNA and an oxygen/water environment are necessarily universal³¹. Xenobiologists, however, would consider oxygen/water to be useful but not essential, carbon-based molecules to be common but not indispensable, and DNA as a strong candidate for a parochial feature that is unlikely to be repeated elsewhere. In contrast, the dual interpretation of DNA as 'instructions' to be carried out and 'information' to be copied, predicted by von Neumann³² on mathematical grounds just before Crick and Watson discovered the structure of DNA, is likely to be a universal. Many aliens will therefore have their own

kind of genetics, because genetics is a useful general trick. But alien genetics might be based on substrates other than DNA. We already know that the double-helix configuration of DNA is only one of many that are possible³³ and that additional artificial bases (now more than twenty) can be included in DNA³⁴. It also seems plausible that synthetic transfer RNAs could be constructed to change the genetic code and even to introduce new amino acids³⁵. Most standard DNA chemistry is parochial, and aliens will not possess it.

Extelligence

A key question for xenobiology is the status of intelligence. Is intelligence a universal? The answer is unclear. Human-level intelligence has arisen only once on Earth, so by normal criteria it ought to be counted as a parochial. On the other hand, intelligence not so different from our own can be found in the great apes, cetaceans and the octopus. Pigs are excellent at video games, parrots have a surprisingly good grasp of linguistics³⁶, and even sticklebacks and mantis shrimps can solve problems. Intelligence looks like it should be a universal because it seems to offer major evolutionary advantages, irrespective of context.

However, the most important ingredient for sentient, technically competent aliens is not intelligence, but a property we have elsewhere called 'extelligence'³⁰. This is the contextual analogue of individual intelligence. Humanity's assumption of global dominance is a tale of extelligence: language, permanent archives of information such as books, and communication in all its technological forms. When compared with most forms of life, our intelligence is only marginally greater than that of chimpanzees: it is our extelligence that has driven our cultural growth and technology. Human extelligence is far more powerful than any individual, but we can all contribute to it, draw on it and exploit it.

On the existing evidence, extelligence may also be a parochial. But again, it looks like such a useful generalized trick that we might be tempted to think of it as a universal. Technologically advanced aliens will, by definition, possess extelligence as well as intelligence. This is where some intelligent species on Earth seem deficient. Dolphins, for example, are able to communicate with one another, but do not appear to be extelligent — we see no dolphin technology. It remains possible that signs of dolphin technology exist but in a form too alien for us to recognize, but we consider this unlikely at present.

Unearthly habitats

Life is a universal, so it will evolve in any habitat that supports the required complexity of organization. We cannot, as yet, define those properties of habitats necessary to support



life with the required degree of generalization, but it is likely that our familiar water/oxygen planet is only one of many possibilities. Science fiction has explored many others, including the surfaces of other planets and asteroids, the atmospheres of gas giants, stellar interiors, interstellar space, molecular clouds, and even the surfaces of neutron stars. Some of these locations, conventionally regarded as passive environments, such as stars and molecular clouds, have occasionally been depicted as life-forms in their own right. In fact, it is difficult to imagine a habitat that could not support a suitable form of life. Anywhere that physical matter can exist, and that offers a rich enough energy substrate, can in principle harbour highly organized processes carried out using matter and energy of the same kind. As far as we are concerned, that is alien life. (We modestly propose our own effort³⁷ as an exploration of the diversity of life when treated as a universal, free from the confines of terrestrial parochiality.)

Where are they, then?

A balloon-like creature floating in the atmosphere of Jupiter would probably regard the terrestrial environment as lethally unattractive. Most aliens would not wish to visit Earth at all, any more than we would care for a ramble across the surface of a neutron star, or to live, as do some extremophiles, in boiling water. We might suppose that the aliens least disinclined to visit us are those who have evolved in an Earth-like habitat, and such habitats might comprise an unknowably small subset of all possible life-supporting habitats. The chances that such aliens exist within 1,000 light years of us at the present time is small. There are plenty of places to visit: why Earth? However, non-humanoid aliens might be keeping a cold, unsympathetic eye on us for their own scientific purposes, writing yet another small footnote in their xenobiology texts.

But if they are here, they will not be easy to spot. As discussed above, they are unlikely to do anything as obvious as abduct gullible readers of supermarket magazines. It is likely that they will possess technology that to us would appear incomprehensible, in accordance with Clarke's dictum³⁸ that "Any

sufficiently advanced technology is indistinguishable from magic." (Or, in Benford's restatement, "Any technology distinguishable from magic is insufficiently advanced."³⁹) Aliens would not look like the canonical Little Green Men. They might look exactly like people. Or cats. Or houseflies. Or they are invisible, or lurking just outside our space-time continuum along a fifth dimension, observing our insides like The Sphere in Flatland observing A. Square⁴⁰. Or they are concealed inside atoms. Or they exist only in the gaps when human perceptual systems are in their refractory phase and unable to observe them.

We think it most likely (and less paranoid) to assume that they are not here at all — for reasons of alien extelligence rather than non-existence. Why run the risk of travelling to exotic places when you can put on a headset and walk through Virtual Venice or Artificial Africa? When VR becomes as real as RealR, an actual visit might seem bothersome, expensive, unsafe and even boring.

We can see the germ of this introspective trend within humanity, so far the only extelligent species we know. More than thirty years ago we landed on the Moon. Our last visit was in 1972, and we no longer have a ready capability to land there. A low-Earth-orbit space station is laboriously taking shape, amid little real enthusiasm. We talk of future manned expeditions to Mars, but a projected unmanned probe to Pluto has been cancelled. The question is not about whether aliens have visited us, and if so, why they aren't here. The important question is why we have not ventured further into space. It would be sad if it turns out that the inability (or reluctance) of an extelligent species to leave home turns out to be a universal. □

1. Marcy, G. W. & Butler, R. P. Detection of extrasolar giant planets. *Annu. Rev. Astron. Astrophys.* **36**, 57–97 (1998).
2. Carr, M. H. *et al.* Evidence for a subsurface ocean on Europa. *Nature* **391**, 363–365 (1998).
3. Chapman, C. R. Probing Europa's third dimension. *Science* **283**, 338–339 (1999).
4. Chyba, C. F. Energy for microbial life on Europa. *Nature* **403**, 381–382 (2000).
5. Kauffman, S. A. *The Origins of Order* (Oxford Univ. Press, Oxford, 1993).
6. Zuckerman, B. & Hart, M. H. (eds) *Extraterrestrials, Where are They?* (Cambridge Univ. Press, Cambridge, 1995).

7. Crawford, I. Where are they? *Sci. Am.* **283**, 38–43 (2000).
8. Adams, D. *The Hitchhiker's Guide to the Galaxy* (Pan Books, London, 1979).
9. Brin, D. *The Uplift War* (Bantam, London, 1987).
10. Fogg, M. J. Temporal aspects of the interaction among the first galactic civilisations: the Interdict Hypothesis. *Icarus* **69**, 370–384 (1987).
11. Benford, G. *Great Sky River* (Gollancz, London, 1988).
12. Cohen, J. & Stewart, I. *The Collapse of Chaos* (Viking, New York, 1994).
13. Pratchett, T., Stewart, I. & Cohen, J. *The Science of Discworld* (Ebury, London, 1999).
14. Artsutanov, Y. N. Komsomolskaya Pravda 31 July (1960).
15. Isaacs, J. D., Bradner, H. & Backus, G. E. Satellite elongation into a true 'sky-hook'. *Science* **151**, 682–683 (1966).
16. Collar, A. R. & Flower, J. W. A (relatively) low altitude 24-hour satellite. *J. Br. Interplanet. Soc.* **22**, 442–457 (1969).
17. Gee, H. *In Search of Deep Time* (Free Press, New York, 1999).
18. SETI Institute Perspectives (SETI Institute, Mountain View, CA, 1999).
19. Tough, A. Fresh SETI strategies. *J. Br. Interplanet. Soc.* **52**, 286–289 (1999).
20. Lachmann, M., Newman, M. E. J. & Moore, C. The physical limits of communication. Working paper 99-07-054, Santa Fe Institute <<http://www.santafe.edu/sfi/publications/Working-Papers/99-07-054.pdf>> (2000).
21. Keller, E. F. *The Century of the Gene* (Harvard Univ. Press, Harvard, 2000).
22. Auyang, S. Y. *Foundations of Complex-System Theories* (Cambridge Univ. Press, Cambridge, 1998).
23. White, R. H. Hydrolytic stability of biomolecules at high temperatures and its implication for life at 250 °C. *Nature* **310**, 430–432 (1984).
24. Russell, M. J., Daia, D. E. & Hall, A. J. in *Thermophiles, the Keys to Molecular Evolution and the Origin of Life* (eds Adams, M. W. W., Ljungdahl, L. G. & Wiegel, J.) 77–126 (Taylor and Francis, Washington, 1998).
25. Stetter, K. O. in *Extremophiles: Microbial Life in Extreme Environments* (eds Horikoshi, K. & Grant, W. D.) 1–24 (Wiley-Liss, New York, 1998).
26. Friedmann, E. I. Endolithic microorganisms in the Antarctic cold desert. *Science* **215**, 1045–1053 (1982).
27. Gold, T. The deep, hot biosphere. *Proc. Natl Acad. Sci. USA* **89**, 6045–6049 (1992).
28. Anderson, R. T., Chapelle, F. H. & Lovley, D. R. Evidence against hydrogen-based microbial systems in basalt aquifers. *Science* **281**, 976–977 (1998).
29. Huber, C. & Wächtershäuser, G. Peptides by activation of amino acids with CO on (Ni,Fe)S surfaces: implications for the origin of life. *Science* **281**, 670–672 (1998).
30. Stewart, I. & Cohen, J. *Figments of Reality* (Cambridge Univ. Press, Cambridge, 1997).
31. Ward, P. D. & Brownlee, D. *Rare Earth: Why Complex Life is Uncommon in the Universe* (Copernicus Books, New York, 2000).
32. von Neumann, J. *Theory of Self-Reproducing Automata* (Univ. Illinois Press, Urbana, IL, 1966).
33. Sinden, R. S. *DNA Structure and Function* (Academic, New York, 1994).
34. Schroppe, M. Expanding life's alphabet. *New Sci.* 8 April, 12 (2000).
35. Atkins, J. F. & Gesteland, R. F. The twenty-first amino acid. *Nature* **407**, 463–464 (2000).
36. Pepperberg, I. M. *The Alex Studies: Cognitive and Communicative Abilities of Grey Parrots* (Harvard Univ. Press, Cambridge, MA, 1999).
37. Cohen, J. & Stewart, I. *Wheeler's* (Warner Aspect, New York, 2000).
38. Clarke, A. C. *Profiles of the Future* (Heinemann, London, 1962).
39. Benford, G. *Foundation's Fear* (Orbit, London, 1998).
40. Abbott, E. A. *Flatland: A Romance of Many Dimensions* (Seeley, London, 1884).

Senescence and cell death

Recent launches include kits to detect senescence in cells, and apoptosis.

CEQ DTCS

Beckman Coulter www.beckmancoulter.com

Rapid set-up for DNA sequencing

The CEQ DTCS Quick Start kit for simplified dye terminator cycle sequencing reaction set-up features a master mix of components that reduces the number of pipetting steps from 10 to four while using larger transfer volumes to reduce pipetting error. The kit works with the CEQ 2000XL DNA sequencing system and can be used with the Biomek FX and Biomek 2000 automated liquid handling systems.

Reader Service No. 100

APOPercentage

Biocolor www.biocolor.co.uk

Rapid apoptosis quantification

The APOPercentage assay kit detects and measures apoptosis in mammalian, anchorage-dependent cells. Analysis can be conducted with a conventional inverted microscope for cell counting, checking cell growth and confluence. The dye contained within the labelled cells and 'blebs' is then released into solution and quantified using a colour filter microplate colorimeter. Assay time is one hour and has been optimized for 96-well plate format; necrotic cells remain unlabelled. APOPercentage is particularly recommended for screening multiple test samples.

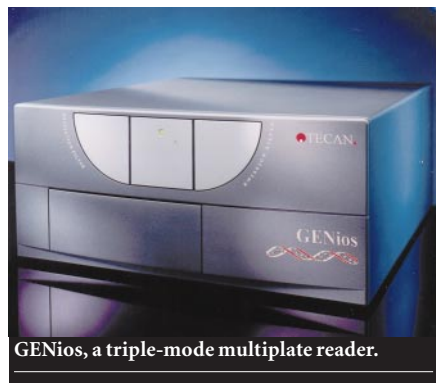
Reader Service No. 101

GENios

Tecan www.tecan.com

Versatile microplate reader

GENios measures fluorescence, absorbance and luminescence and can be used in a range of applications, including expression studies (GFP and luciferase-based), binding studies, kinetic enzyme investigations, toxicology and ELISAs. In cell-based assays looking at cytotoxicity, proliferation or apoptosis, the reader can detect multiple labels per well and can be



GENios, a triple-mode multiplate reader.



CEQ 2000XL sequences 600 nucleotides in less than 50 minutes.

switched from top to bottom reading with the click of a mouse. It can be used with all microplate formats (from six to 384 wells), PCR tubes and cuvettes.

Reader Service No. 102

Senescence detection kit

Trevigen www.trevigen.com

Check the cell by date

Senescent cells are at the end of their replicative life and are characterized by a loss of response to proliferating or apoptotic stimuli. Metabolic activities are maintained in senescent cells, but they present a flattened and enlarged morphology. Normal ageing, introduction of oncogenes or treatment with some biological compounds are examples of senescence inducers. This kit, described as the first commercial kit developed for detection of senescence in cells, uses a chromogenic substrate that forms a blue precipitate at the sites of senescence-associated β -galactosidase. The kit comes with positive control reagents, counterstaining reagent and detailed instructions.

Reader Service No. 103

ExpressDirect

Pierce www.piercenet.com

mRNA capture and RT system

A complete system for cell lysis, mRNA capture and reverse transcription for RT-PCR, this kit uses a two-step lysis procedure that selectively disrupts only the outer cell membrane so that processed mRNA — and not contaminating genomic DNA — is released

from the starting cell or tissue sample. The mRNA is then captured directly from the cell lysate by hybridization to oligo(dT) that is covalently linked onto the wells of a 0.2-ml, 96-well PCR plate. First-strand cDNA synthesis of the captured mRNA is performed in the same tube. The flexible microtube plate format allows 1–96 reactions to be performed. A product instruction book can be downloaded from the company's website.

Reader Service No. 104

S.N.A.P.

Invitrogen www.invitrogen.com

UV-free gel purification kit

Traditional methods of DNA visualization involve staining agarose gels with ethidium bromide under UV light, which nicks and damages DNA and reduces cloning efficiency. This kit uses a crystal violet staining reagent to allow detection of DNA in agarose gels under ambient light. After visualization, DNA up to 20kb can be gel-isolated using a simple spin column format. Eluted DNA is ready for use in transfection, sequencing or cloning.

Reader Service No. 105

PCR tubes

Eppendorf www.eppendorf.com

Sealed with a hinge

These 0.2-ml thin-PCR tubes with contamination shields now boast an optimized sealing mechanism consisting of a hinge that allows the lid to lock into position at 90°, relieving stress on the sealing lip and eliminating the risk



Eppendorf's anti-contamination PCR tube.

of evaporation during the PCR. Extra labelling space is provided on the body of the tube, in addition to that on the frosted lid. Accessories include a workplate in the microtitre plate format that can be used to insert the tubes in a 96-well thermoblock. In combination with a frame, it also functions as a tube rack. The tubes are available individually or in strips of five or eight.

Reader Service No. 106

Quantitative analysis of DNA adducts

Micromass www.micromass.co.uk

A new application note

'The Quantitative Analysis of DNA Adducts Using MicroHPLC/Platform ICP-MS' (application note AN306) describes the use of inductively coupled plasma-mass spectrometry (ICP-MS) to analyse DNA adducts of xenobiotic compounds thought to play a key role in human cancer induction. The note covers atom-specific analysis methods such as ICP-MS in the quantification of phosphorus using generic standard solutions. The change in solvent composition during a gradient elution HPLC run presents a challenge to ICP-MS analysis. Surface tension, viscosity, density and a decreasing desolvation can all cause signal depression. However, use of a micro-concentric nebulizer (MCN) with desolvation as a coupling interface between the HPLC and the ICP-MS stabilizes signal intensity over a wide range of solvent compositions.

Reader Service No. 107

Alu probe

InnoGenex www.innogenex.com

Detecting primate-specific sequences in tissues

The *Alu* repetitive elements, unique to primates, are interspersed within the human genome with an average spacing of 4kb. Retro-transposal integrations of *Alu* and L1 sequences into biologically important genes

appear to be key factors in some human diseases. Although the human *Alu* family, and related families, originates from 7SL RNA, all other SINEs originate from tRNA. The InnoGenex *Alu* probe will detect *Alu* repeat sequences found in primate chromosomes in formalin-fixed, paraffin-embedded tissue sections by *in situ* hybridization. The *Alu* probe is a mixture of synthetic oligonucleotides labelled with five fluorescein moieties at the 5' end and is designed for efficient detection using anti-fluorescein antibodies. The probe can be used to localize human tissues in chimeric animals. Rodents with human cancer implants can be stained with the *Alu* probe to determine the extent of cancer proliferation.

Reader Service No. 108

Gibco BRL OligofectAMINE

Life Technologies www.lifetech.com

Optimized for antisense applications

OligofectAMINE reagent can be used for the delivery of functional antisense oligonucleotides into a wide variety of cultured eukaryotic cells, including CHO, K562, HEK293, HeLaS3 and HeLa. It is effective with small amounts of oligonucleotide and works with nuclear and cytoplasmic targets. Non-toxic and easy to use, the reagent is particularly suited to high-throughput applications.

Reader Service No. 109

Shrimp alkaline phosphatase

Promega www.promega.com

This SAP can't stand the heat

Alkaline phosphatases are used for the dephosphorylation of 5'-phosphorylated DNA or RNA ends. Dephosphorylated DNA can either be re-ligated in cloning experiments or labelled with ³²P-γATP and T4 polynucleotide kinase. Dephosphorylation prevents the re-ligation of linearized plasmid DNA in cloning experiments. Unlike calf intestinal alkaline phosphatase, shrimp alkaline phosphatase (SAP) is completely and irreversibly inactivated by heat treatment for 15 minutes at 65°C.

Reader Service No. 110

ProxiPlate-384 F

Packard Instrument

www.packardinstrument.com

Close call

Suggested for use with the company's Homogeneous Time-Resolved Fluorescence (HTRF) technology measured in the Discovery microplate analyser, the black ProxiPlate-384 F is manufactured from a low-fluorescence resin and is suitable for fluorescence applications where sterility is not required. It has a shallow well with a total volume of 30 μl. Samples are thus positioned close to the instrument's detectors, produc-

ing enhanced light signal output, better sensitivity and effective miniaturization with reduced reagent costs.

Reader Service No. 111

arraySCOUT 2.0

LION bioscience www.lionbioscience.com

Upgraded system for microarray data analysis

Improvements to arraySCOUT 2.0 include the ability to interface with any database system containing 'raw' expression data, a powerful set of new clustering and visualization tools and an open application programming interface to allow users to add their proprietary analysis methods. arraySCOUT 2.0 is fully integrated with two upcoming software modules for the analysis of biological pathways and molecular interaction data. For more information visit the company web site, or contact arraySCOUT@lionbioscience.com.

Reader Service No. 112

These notes are compiled in the Nature office from information provided by the manufacturers.